STOCHASTIC M-ESTIMATORS: CONTROLLING ACCURACY-COST TRADEOFFS IN MACHINE LEARNING

A Thesis
Presented to
The Academic Faculty

by

Joshua V. Dillon

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in
Computational Science & Engineering

College of Computing
Georgia Institute of Technology
December 2011

STOCHASTIC M-ESTIMATORS: CONTROLLING ACCURACY-COST TRADEOFFS IN MACHINE LEARNING

Approved by:

Professor Guy Lebanon, Advisor
College of Computing
Georgia Institute of Technology

Professor Alexander Gray
College of Computing
Georgia Institute of Technology

Professor Mark Borodovsky
College of Computing
Georgia Institute of Technology

Professor Hongyuan Zha
College of Computing
Georgia Institute of Technology

Doctor Kevyn Collins-Thompson
Microsoft Research

Date Approved: November 14, 2011

To Prerna.

ACKNOWLEDGEMENTS

To my friends and colleagues. I would not have survived my first day of grad school–let alone see it to completion–without your friendship and words of encouragement. Thank you for making this journey so much fun!

To my family: Mom, Dad, Greg & Jill. You have always provided me with a refuge from my fears, my obsessions, and myself. With much love, I thank you.

To my mentor, Kevyn. It has always been clear to me how much you want me to succeed. Thanks for all the opportunities you unlocked and the good times we had on the way.

To my advisor, Guy. I cannot possibly express the depth my gratitude. True to academic tradition, you opened my eyes to a world I'd otherwise never known. This is good. But not great. For that, I look to you: brilliant, creative, humble, kind, quick to laugh, and a great friend. Thank you.

And finally to my Prerna, Prerna. You fixed me. Because of you I now walk tall. Because of you, I now walk with purpose.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

CoNLL	Conference on Computational Natural Language Learning.
POS	Part-of-speech (tagging).
WSJ	Wall Street Journal corpus.
BFGS	Broyden-Fletcher-Goldfarb-Shanno numerical optimization method.
EM	Expectation Maximization.
FL	Full likelihood.
FLOP	Floating point operation.
MCMC	Markov chain Monte Carlo.
MSE	Mean Squared Error.
NLP	Natural Language Processing.
PAC	Probably Approximately Correct.
PL-k	Pseudo Likelihood, order k (PL when $k = 1$).
RV	random variable.
SSL	Semi-supervised Learning.
BM	Boltzmann Machine.
CRF	Conditional Random Field.
HMM	Hidden Markov Model.
MRF	Markov Random Field, i.e., HMM, CRF.
$D(p\|\|q)$	KL-Divergence, $D(p\|\|q) = H(p,q) - H(p)$.
$H(p)$	Entropy, $H(p) = H(p,p)$.
$H(p,q)$	Cross-entropy, $-\int_{\mathcal{X}} p(x) \log q(x) dx$.
PSD	positive semi-definite matrix, i.e., $x^* M x \geq 0$ for all $x \in \mathbb{C}^n$.
$S_\theta(A_j, B_j)$	Score; $S_\theta(A_j, B_j) = \log p_\theta(X_{A_j} \| X_{B_j})$.
$\hat{\theta}_n^{\mathsf{ml}}$	Maximum likelihood estimator (MLE).
$\hat{\theta}_n^{\mathsf{mpl}}$	Maximum pseudo likelihood estimator (MPLE).

$\hat{\theta}_n^{\mathsf{msl}}$	Maximum stochastic composite likelihood estimator (MSCLE).
(A, B)	m-Pair, partitioning of dimensions of X, i.e., $\{1 \dots m\} = A \cup B$ and $A \cap B = \emptyset \neq A$.
SCL	Stochastic Composite Likelihood.
AVar	Asymptotic Variance.
β	Vector of non-negative m-function weights.
D	Data, collection of iid samples X, (X, Y), or (W, X, Y).
λ	Vector of m-function selection probabilities.
$M_n(\theta; D)$	m-Estimator criterion, an average of (non-random) m-functions.
$m_\theta(X)$	m-Function, a non-random known function of a data sample X.
P	Distribution of nature, i.e., the true data generating process.
Υ	Sum of variances (component gradient score).
Σ	Variance of sum (component gradient score).
$\widetilde{M}_n(\theta; D)$	Stochastic m-estimator criterion, an average of stochastic m-functions.
$\widetilde{m}_\theta(X, Z)$	Stochastic m-function, a cross-correlation of activator and m-functions.
θ	Model parameter identifying a member of $\{p_\theta : \theta \in \Theta\}$.
θ_0	Model parameter which is "closest" to distribution of nature P.
X	Evidence RV, sample drawn from a distribution of nature, P.
Y	Observed RV, drawn from a distribution of nature, $P(\cdot\|X)$.
Z	m-Function activator, binary random variable.
$\chi_j(Y)$	Instantiated labeling policy, a subset of labels of Y.
$\wp(Y)$	Labeling Policy RV, a random mapping of label sequences Y to subsets of labels $\chi_j(Y)$ with proabbility λ_j.

SUMMARY

m-Estimation represents a broad class of estimators, including least-squares and maximum likelihood, and is a widely used tool for statistical inference. Its successful application however, often requires negotiating physical resources for desired levels of accuracy. These limiting factors, which we abstractly refer as costs, may be computational, such as time-limited cluster access for parameter learning, or they may be financial, such as purchasing human-labeled training data under a fixed budget. This thesis explores these *accuracy-cost tradeoffs* by proposing a family of estimators that maximizes a stochastic variation of the traditional m-estimator.

Such "stochastic m-estimators" (SMEs) are constructed by stitching together different m-estimators, at random. Each such instantiation resolves the accuracy-cost tradeoff differently, and taken together they span a continuous spectrum of accuracy-cost tradeoff resolutions. We prove the consistency of the estimators and provide formulas for their asymptotic variance and statistical robustness. We also assess their cost for two concerns typical to machine learning: computational complexity and labeling expense.

For the sake of concreteness, we discuss experimental results in the context of a variety of discriminative and generative Markov random fields, including Boltzmann machines, conditional random fields, model mixtures, etc. The theoretical and experimental studies demonstrate the effectiveness of the estimators when computational resources are insufficient or when obtaining additional labeled samples is necessary. We also demonstrate that in some cases the stochastic m-estimator is associated with robustness thereby increasing its statistical accuracy and representing a win-win.

CHAPTER I

INTRODUCTION

To establish some basic context, we begin by informally introducing a statistical technique known as m-estimation and the key reasoning behind its widespread employ in the fields of statistics and machine learning. We then briefly describe the primary contribution of this dissertation, the stochastic m-estimator, as an extension of the standard m-estimator and describe the practical and analytical benefits afforded by this seemingly simple modification. We conclude this chapter with an organizational overview of the dissertation and a synopses of subsequent chapters.

Throughout this dissertation, we are fundamentally concerned with finding the index θ which identifies a member of a family of known, non-random functions $\{m_\theta : \theta \in \Theta\}$. The function should best characterize n, m-dimensional random variates,

$$D_n = (x^{(1)}, \ldots, x^{(n)}), \quad x^{(i)} \in \mathcal{X} \subset \mathbb{R}^m, \tag{1}$$

which are assumed independent and identically sampled according to law or distribution P. We call D_n the dataset and by the interchangeability implied by the iid assumption, we denote the random set by $\{X^{(i)}\}_1^n$. Following standard conventions, we denote random variables by upper case and random variates (instantiations) by lower case.

A popular method for making precise the notion of "best characterize," is to specify each $m_\theta(x)$ such that it reflects some indication of utility or negative loss with respect to a particular instance x. Under this analogy, the point θ may be regarding not just as identifying a particular member of the family, but as perhaps possessing its own intrinsic meaning, for example, it is often a parameter or knob which controls how dimensions of $\mathcal{X}$ relate to one another and how those relationships should be

combined to indicate utility. Reasonably, one then seeks the θ which maximizes this measure, i.e.,

$$M_n(\theta; D) = \frac{1}{n} \sum_{i=1}^{n} m_\theta(x^{(i)}). \tag{2}$$

Since the functions are non-random, the sequence $\{m_\theta(x^{(i)})\}_1^n$ remains iid, and we may expect to exploit classical statistical techniques in its analysis. The value(s) for which $M_n(\theta)$ attains its maximum is denoted $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$. Being a function of random values, $\hat{\theta}_n$ is itself random variable; we refer to a particular realization as the estimate or estimator.

Estimators which maximize criteria of the form of (2) are known as m-estimators, and represent a broad class of point estimation techniques. They are a standard tool in statistical inference and machine learning, with classic examples being least-squares estimators and the maximum likelihood estimator (MLE). We refer to a particular function of the data as an m-function, for example, the MLE is defined as $m_\theta(X) = \log p_\theta(X)$ where p_θ is a probability function parameterized by θ. Like the MLE, m-estimators enjoy several favorable statistical qualities, such as consistency[1] and asymptotic Normality. Unlike the MLE though, the m-functions, may not directly represent a probability density function. Hence m-estimators are not fully parametric and in this sense more general than the MLE.

Traditionally, m-estimators were motivated by the need for estimators which are immune to small departures in model assumptions. However, since many standard point estimators are also m-estimators, they provide a broad framework for exploring the statistical properties of a variety of techniques. Although there are many reasons for studying these estimators, we focus on three main aspects: (i) tailoring m-functions to satisfy the modeler's requirements, i.e., computational and asymptotic efficiency,[2] (ii) robustness or resistance to deviations from assumptions (vs., say, the

[1] That is, loosely speaking, it converges in probability to the true value of the estimator.

[2] The extent which it is estimated in some "best possible" manner, under large data assumptions.

MLE), and, (*iii*) they can be analyzed using techniques that do not assume the true model is within the assumed parametric family.

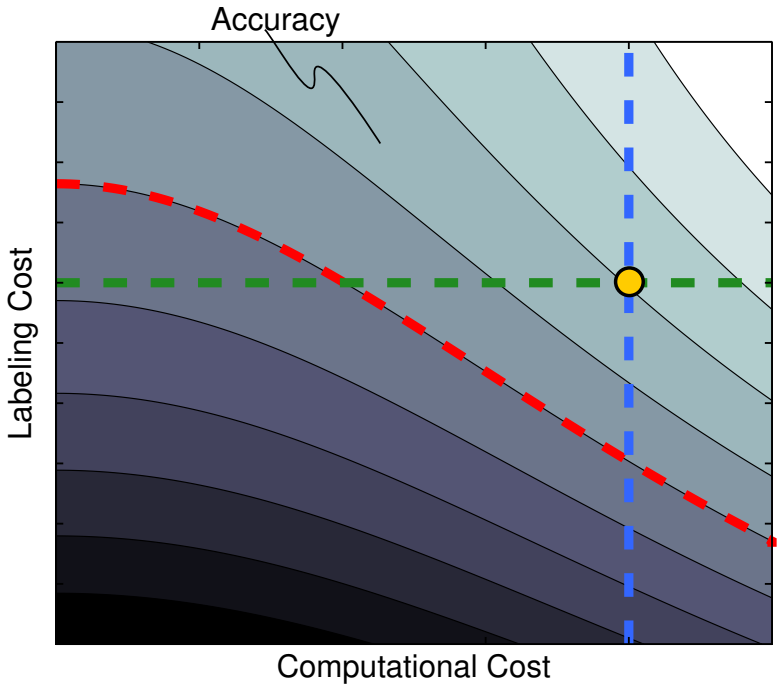

Figure 1: Hypothetical tradeoff between accuracy (contours; decrease in darkness), computational cost (x-axis), and labeling cost (y-axis). The vertical line represents realizable accuracies given a fixed computational budget while the horizontal corresponds to fixed labeling budget. The curve represents all combinations of computational and labeling costs that achieve a certain accuracy.

Unfortunately many m-estimators are not practical, despite possessing many favorable qualities. For example, although statistically efficient, log-likelihood m-estimators for many different models are of limited practical use due the *computational cost* of evaluating the m-function. Conversely, the pseudo log-likelihood m-estimator, a "relaxation" of the log-likelihood (made precise in Chapter 3), trades computational costs for reduced statistical efficiency, that is, it pays a higher *statistical cost*. Budgeted learning scenarios, i.e., active learning and generative semi-supervised learning, represent another common example. Here, the evaluation of each m-function has a *labeling cost* (due to purchasing labeled examples) which must be negotiated to achieve accuracy requirements.

Figure 1 depicts these notions as hypothetical example; in this case it is a three-way tradeoff between accuracy, computational cost, and labeling cost. The contours are logarithmic with darker shades indicating decreased accuracy. For this illustration, we assume the true model is computationally challenging to fit and there exists a reasonable simplification scheme. Such a situation commonly occurs when modelers intentionally assert false independencies to render parameter learning tractable. Labeling costs correspond to purchasing human-labeled training samples; we may reasonably imagine paying people to identifying man-made structures in images or nouns in sentence.[3] Examining this space of tradeoffs reveals many of the usual regimes present in data modeling:

1. *Fixed computation*: accuracy increases with additional labeled samples. This regime is often employed as a theoretical framework for the analysis of estimators but also represents active learning or dataset construction.

2. *Fixed labels*: accuracy increases with additional computation. This regime characterizes the analysis of approximate inference and has numerous practical implementations, such as Monte Carlo techniques.

3. *Fixed labels & computation*: accuracy is fixed at single point in the tradeoff space. This regime includes fixed dataset, fixed inference problems and represents many traditional machine learning tasks.

4. *Fixed accuracy*: combinations of computational and labeling costs with the same accuracy. This regime represents a holistic view of costs and their relation to accuracy.

The main trend of Figure 1 is typical of many, if not most learning problems; higher accuracy necessitates a large set of labeled samples and increasingly precise

[3]Obviously these approaches would not result in such smooth tradeoffs; this is in fact a key contribution of this dissertation.

model fitting. Perhaps more interesting, are the variety of ways we achieve the same level of accuracy (red curve) and how we can exploit the interplay between costs. It seems that at lower accuracies, it may be more beneficial to invest in obtaining larger training sets rather than refining the approximation scheme. Following this trend from bottom-left to top-right, we note that at some point the two costs decouple and we can obtain the same amount of improvement from either. However this phenomenon is not without drawback–it takes increasingly more computation and/or labels to achieve the same levels of improvement in accuracy, i.e., this region represents diminishing return.

Although the previous issues have been extensively studied in one form or another, we are unaware of any preexisting, holistic treatment capable of quantifying both statistical performance and the associated costs, while simultaneously providing practical algorithms for smoothly realizing the continuum of tradeoffs. This dissertation offers such a framework by presenting a family of estimators which maximize a stochastic variation of the traditional m-estimator.

1.1 The Stochastic m-Estimator

The standard application of m-estimators entails selecting a fixed m-function a priori, either from a variety of well-studied functions or a proposed alternative. Indeed assessing which m-function to employ is a well-studied problem, however, operationally speaking the choice remains a one-time procedure. In this regard there is a certain lack of flexibility associated with the m-estimator framework. There is no ability for some m-functions to be selected more frequently than others, as resource budgets permit at that particular instant. For example, available computational resources may allow the computation of the full log-likelihood for 20% of the samples, and the pseudo log-likelihood for the remaining 80%.

This dissertation addresses this shortcoming via a stochastic extension of traditional m-estimators. Loosely speaking, stochastic m-estimators (SME) stitch together different m-estimators, at random. In doing so, each such instantiation resolves the accuracy-cost tradeoff differently, and taken together they span a continuous spectrum of accuracy-cost tradeoff resolutions. Although we formally develop SMEs subsequently, the general form remains much like (2), with the exception of an additional random argument Z, i.e.,

$$\widetilde{M}_n(\theta\,; D, Z) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{m}_\theta(X^{(i)}, Z^{(i)}), \quad \text{with} \tag{3}$$

$$\widetilde{m}_\theta(X, Z) = \sum_{j=1}^{k} \beta_j Z_j m_{j\theta}(X), \tag{4}$$

for some non-negative weighting scheme $\beta \in \mathbb{R}_+^k$, which we often drop from the notation for brevity. The random variables $Z = (Z^{(1)}, \ldots, Z^{(n)})$ are each binary k-dimensional vectors and sampled independently from some density f_λ. Note that in general $Z_r^{(i)}$ may not be independent from $Z_{-r}^{(i)}$ but it is independent of $Z_:^{(j)}$ and $X^{(i)}$. Where convenient we intermingle the signal processing notation $Z_{2:7}$, which indicates a subset of the dimensions two through seven and set notation Z_{-4} which indicates all dimensions except four. Since $Z_j^{(i)}$ is binary, the function $\widetilde{m}_\theta(X)$ exists as a randomly affine combination of k possibly evaluated m-functions, $m_{j\theta}(X)$.

Under this construction, $\widetilde{M}_n$ is tunable through three mechanisms:

1. The m-functions themselves.

2. The weights β.

3. The density f_λ associated with Z.

We refer to λ as the selection "policy." The policy λ and the characteristics of $m_{j\theta}$ dictate where in the tradeoff space, i.e., Figure 1, the estimator should reside. Choosing a policy $\lambda \in \Lambda$ should not be regarding as selecting a hyperparameter but

rather a tunable knob that represents a commitment of the modeler's resources to his or her task. Different policies may emphasize or de-emphasize different m-functions depending on their characteristics. Unlike the policy, the weights β are properly regarded as nuisance parameters and control the statistical efficiency of the SME. This issue is given full treatment in subsequent chapters along with a simple heuristic for its selection.

We note that the criterion (3) cannot be re-expressed as an m-estimator despite being a sum of sums. The sum in (3) is over independent terms while the terms of (4) are clearly not independent as they are evaluated at the same X.

1.2 Organization

This work develops the stochastic m-estimator for two fundamental issues of many statistical machine learning tasks: computational costs of parameter learning and labeling costs of dataset construction. We study these tradeoffs and show that SMEs result in practical algorithms that can smoothly negotiate them.

The remainder of this dissertation is organized as follows. In Chapter 3, we explore the computational tradeoffs implicit to parameter learning. For the sake of concreteness, we limit this discussion to Markov random fields. In Chapter 5, we explore the labeling tradeoffs present in budgeted learning. Here again we limit the analysis to generative semi-supervised learning scenarios for the sake of simplicity. We discuss the relevant related work separately in each chapter. Since we tailor each SME to the particular accuracy-cost tradeoff, we also separately explore its statistical properties.

1.2.1 Computational Costs

Chapter 3 describes the use of stochastic m-estimators in situations where the computation of the MLE is intractable. In contrast to many previously proposed approximate estimators, this estimator is statistically consistent and admits a precise

quantification of both computational complexity and statistical accuracy through analyzing its asymptotic variance. Due to the continuous parameterization of the estimator family, we obtain an effective framework for achieving a predefined problem-specific balance between computational tractability and statistical accuracy. We also demonstrate that in some cases reduced computational complexity may in fact act as a regularizer, increasing robustness and therefore accomplishing both reduced computation and increased accuracy. This "win-win" situation conflicts with the conventional wisdom stating that moving from the MLE to pseudo likelihood and other related estimators result in a computational win but a statistical loss. Nevertheless we show that this occurs in some practical situations.

We discuss experimental results in the context of Boltzmann machines and conditional random fields. The theoretical and experimental studies demonstrate the effectiveness of the estimators when the computational resources are insufficient. They also demonstrate that in some cases reduced computational complexity is associated with robustness thereby increasing statistical accuracy.

1.2.2 Labeling Costs

Chapter 5 examines the use of the stochastic m-estimator for budgeted learning scenarios. Specifically, we explore generative semi-supervised learning. SSL has emerged as a popular framework for improving modeling accuracy while controlling labeling cost. SMEs allow us characterize the asymptotic accuracy of generative semi-supervised learning as a function of number of samples and proportion labeled. The SME framework not only affords the ability to make this characterization for simple scalar labels, i.e., classification, but allows the analysis of multidimensional labels, i.e., structured prediction, thereby permits partial labeling schemes.

By providing analysis for large data, we complement distribution-free approaches by providing an alternative framework to measure the value associated with different

labeling policies and resolve the fundamental question of how much data to label and in what manner. We demonstrate the effectiveness of the SME framework with both simulation studies and real world experiments using naive Bayes for text classification and MRFs and CRFs for structured prediction in NLP.

CHAPTER II

THE STOCHASTIC M-ESTIMATOR

In this chapter we provide a formal development of the stochastic m-estimator (SME). We first define a general functional form of the SME, then we introduce the requisite empirical process theory. Finally we provide proofs of its consistency and asymptotic normality. The techniques that we use to develop the SME can largely be found in [9] and [49].

2.1 Notation

Unless stated otherwise, we will assume we are given a sequence $X_n = \{X^{(i)}\}_{i=1}^n$ of independent and identically distributed random variables with distribution P_X on a measurable space $(\mathcal{X}, \mathcal{A}, \mu)$ where $\mathcal{X} \subseteq \mathbb{R}^m$ is the event space, $\mathcal{A}$ is a σ-algebra, and μ is a σ-finite measure. Measure μ could be a counting measure (when $\mathcal{X}$ is denumerable) or a Lebesgue measure; when it exists, $\mathrm{d}P(x) = p(x)\mathrm{d}\mu(x)$. As is convention, write random variables as upper case and random variates as lower case. We often write $X = X^{(1)}$ when it is unambiguous. Unless stated otherwise, $d(\cdot, \cdot)$ is the Euclidean distance, $I[\cdot]$ the indicator function of some event, and $\delta(y)$ the Dirac delta function (which is 1 if and only if $y = 0$).

In this chapter we use de Finetti notation [41] to simplify the otherwise cumbersome, more standard notation for expectations. We therefore may write the empirical

expectation (average), population average, and empirical process as,

$$P_n f = \frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})$$
$$Pf = \int f \mathrm{d}P$$
$$G_n f = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(x^{(i)}) - Pf) = \sqrt{n}(P_n - P)f.$$

For example, the law of large numbers could be written as $P_n f \to Pf$ almost surely (when Pf exists). The central limit theorem asserts that $G_n f$ is asymptotically normal if $Pf^2 < \infty$.

We use the notation $f : A \mapsto B$ to indicate f is a function with domain A and co-domain B. Often we treat a part of the function's domain as fixed, e.g., $f_\theta(x)$ has a domain consisting of a parameter part θ and a data part x. We use the notation $\theta \mapsto f_\theta(x)$ to indicate that the map under consideration is a function of θ (and fixed for x). We may also write functionals as $f \mapsto f(x)$ which indicates that the domain is $\mathcal{F} = \{f(x) : \mathcal{X} \to \mathbb{R}\}$ and not $\mathcal{X}$. We may also write $K(\theta; x)$ to indicate $\theta \mapsto K(\theta, x)$.

2.2 The Stochastic m-Estimator

The stochastic m-estimator (SME) is a point of maximum of the map,

$$\theta \mapsto P_n \int_{\mathbb{R}} Z(t) m_{\theta,t}(X)\, \mathrm{d}\nu(t), \tag{5}$$

where P_n is the joint empirical distribution of (X, Z), i.e., $p_n(x, z) = P_n \delta_x \delta_z = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x^{(i)})\delta(z - z^{(i)})$. Here $Z(t)$ is a random variable and t an index; we postpone further details for the moment. The function $\nu(t)$ is a σ-finite measure which specifies how the m-functions should be combined. For example, in this thesis it is often the case that $\nu(t)$ is a discrete measure and $Z(t)$ is a binary, length-k random vector. In this case, Equation 5 takes the somewhat simpler form,

$$\theta \mapsto P_n \sum_{j=1}^{k} Z_j m_{\theta,j}(X). \tag{6}$$

The functions $m_{\theta,t} : \mathcal{X} \to \mathbb{R}$ are real-valued maps which are measurable in X for every $\{t : \nu(t) > 0\}$ and $\theta \in \Theta$ where (Θ, d) is an r-dimensional metric space. Individually, each member of the collection $\{m_{\theta,t}(X) : \nu(t) > 0\}$ can be understood as a "deficient" m-function, that is, possessing all the standard properties of an m-function with the exception that $\theta \mapsto P_n Z(t) m_{\theta,t}(X)$ typically has many maxima, even for large n. Collectively, (5) can be understood as specifying a random, composite m-function which is assumed not deficient.

As an example, recall the standard least-squares estimator $m_\theta(X) = \|X - \theta\|_2^2$. A possible SME generalization of this m-estimator is defined by the collection of L_p norms, i.e., $m_{\theta,t}(X) = \|X - \theta\|_t^t$ with $\nu(t) = I(t \in \mathbb{Z}_+)$. Assuming X is a length-m random vector, another possible generalization is to define an SME based on a collection of L_2-normed subsets, i.e., $m_{\theta,t}(X) = \|X_{A_t} - \theta_{A_t}\|_2^2$ where $\nu(\cdot) > 0$ indexes sets $A_t \subseteq \{1 \dots m\}$.[1] Just as with the standard m-estimator, the appropriateness of a particular $\{m_{\theta,t}\}_t$ depends fundamentally upon the task at hand.

Defined on probability space $(\mathcal{Z}, \mathcal{A}_z, P_{\lambda,X})$, the random variable Z controls the relative importance the corresponding $m_{\theta,t}(X)$. The collections $Z = \{Z(t) : \nu(t) > 0\}$ and $\{m_{\theta,t}(X) : \nu(t)\}$ are random processes, thus (5) has the dual-interpretation of being the zero-lag, cross-correlation.[2] Due to this similarity, we often refer to t as a time index and regard $\nu(\cdot)$ as specifying time as discrete or continuous.

The parameter λ of the distribution of $P_{\lambda,X}$ is more a matter of descriptive convenience rather than necessity. Often we refer to a particular λ as a "policy" in order to emphasize its role as a "knob" in selecting particular m-pairs. Procedurally, the process Z is is generated from $P_{\lambda,X^{(i)}}$ for each $X^{(i)}$, and is denoted $Z^{(i)}$. In this fashion the maximization of the SME criterion is deterministic for any given sequence of $\{(X^{(i)}, Z^{(i)})\}_1^n$; this is a matter of practical consequence when designing $\arg\max$

[1] The notation X_{A_t} indicates the sub-vector of X indexed by A_t and is equivalent to $\{X_j : j \in A_t\}$.

[2] For continuous functions, cross-correlation is defined as $(f \star g)(t) \stackrel{\text{def}}{=} \int_{\mathbb{R}} f^*(\tau) g(t+\tau) d\tau$ where f^* denotes the complex conjugate. Since Z is real, (5) is equivalently represented by $P_n[(Z \star m_\theta)(0)]$.

procedures. In this thesis, the random variable $Z(t)$ is typically be independent of X, although strictly speaking, this is not necessary.

2.2.1 Goals

The aim of this chapter is to establish conditions for ensuring that the point-wise maximum of (5), denoted $\hat{\theta}_n$, is equivalent to the maximum of $\theta \mapsto P \int_{\mathbb{R}} Z(t) m_{\theta,t}(X) \, \mathrm{d}\nu(t)$, in the limit of n. In the latter case, we denote the maximizers by Θ_0 with θ_0 being individual members; ideally $|\Theta_0| = 1$ but this is often not so. Convergence of $\theta_0 \to \Theta_0$ shall be understood as $d(\theta, \Theta_0) = \inf\{d(\theta, \theta_0) : \theta_0 \in \Theta_0\} \to 0$ in the appropriate probabilistic sense. Unfortunately, proving the convergence of $\theta_n \to \Theta_0$ as $n \to \infty$ is not as simple as directly applying the Strong Law of Large Numbers; the maximization of the empirical expectation necessitates the development of a uniformly strong law of large numbers. This is precisely the study of empirical process, which we review subsequently.

Before examining conditions which imply $\hat{\theta}_n$ converges to the population maximizer, it is worthwhile to consider how different choices of $P_{\lambda,X}$ result in different limiting points. Although the specification of $P_{\lambda,X}$ is left intentionally ambiguous, we list here a few possible assumptions which are useful to the accuracy/cost tradeoffs developed in subsequent chapters. In the following cases we assume X and Z are independent, i.e., $P_1 = P_X P_\lambda$.

Discrete. When $\nu(t)$ is the discrete measure, we obtain a sum-form SME, i.e., $\theta \mapsto \sum_{t=1}^{k} w_t P_X m_{\theta,t}(X)$ where $w_t = P_\lambda Z_t$ where Z_t is one of the coordinates of Z such that $\nu(t) > 0$.

Independent. When $i \neq j$ implies $Z(i), Z(j)$ are independent, we obtain an SME of the form, $\theta \mapsto \int_{\mathbb{R}} w(t) P_X m_{\theta,t}(X) \, \mathrm{d}\nu(t)$ where $w(t) = P_\lambda^{(t)} Z(t)$. Here $P_\lambda^{(t)}$ denotes the appropriate marginal distribution.

Binary. When $Z(t) \in \{0, 1\}$, we obtain an SME of the form, $\theta \mapsto \int_{\mathbb{R}} w(t) P_X m_{\theta,t}(X) \, \mathrm{d}\nu(t)$ where $w(t) = p_\lambda(Z(t) = 1)$ where p_λ is the appropriate mass function.

Unbiased. Setting $z(t) = 1/p_\lambda(z(t))$ with probability $p_\lambda(z(t)) > 0$ results in an SME of the form $\theta \mapsto \int_{\mathbb{R}} P_X m_{\theta,t}(X) \, \mathrm{d}\nu(t)$.

From these examples it is clear that independence among $Z(i)$ and $Z(j)$ is perhaps only useful for sampling purposes. Likewise discreteness and binary cases are useful when closed form expressions are lacking. More interesting is the ability to remove Z by making it a random variable reciprocal to its probability. We return to this notion in later chapters. In this chapter we show that $X \perp\!\!\!\perp Z$ is a sufficiently strong enough assumption for consistency, that is loosely speaking, a weak kind of unbiasedness.

For the specific distributional forms of P_λ, we often choose either the Multinomial distribution with one draw, or a product of independent Bernoulli distributions. When necessary, we embed a multiplicative constant into each $m_{\theta,t}$ to ensure that all such functions are of comparable scale (rather than embed the constant into Z).

As stated, the aim of this chapter is to show that (5) converges in probability to a point θ_0 which is a maximum of the map,

$$\theta \mapsto P \int Z(t) m_{\theta,t}(X) \, \mathrm{d}\nu(t), \tag{7}$$

where $P = P_X P_{Z|X,\lambda}$. We call this map the population criterion. Since the sample criterion (5) is an average over the observations, we now review the essential tools from empirical processes theory to make this characterization.

2.3 Essential Empirical Process Theory

We now review the fundamental tools from Empirical Process theory necessary for characterizing the asymptotic behavior of the Stochastic m-Estimator. This review is brief and merely for the reader's convenience; proofs and further discussion can be found in the excellent textbook treatment [50].

2.3.1 Stochastic Order Notation

Throughout this thesis, we will find it convenient to denote the convergence of random sequences by the appropriate stochastic order notation.

The notation $o_P(1)$, read "small oh-P-one," denotes a sequence of random vectors that converges to zero in probability. The expression $O_P(1)$, read "big oh-P-one," denotes a sequence bounded in probability. In other words, for a given sequence of random variables R_n,

$$X_n = o_P(R_n) \quad \text{means} \quad X_n = Y_n R_n \quad \text{where} \quad Y_n \xrightarrow{\text{p}} 0,$$

$$X_n = O_P(R_n) \quad \text{means} \quad X_n = Y_n R_n \quad \text{where} \quad Y_n = O_P(1).$$

Here, $X_n \xrightarrow{\text{p}} X$ means for all ε, $PI[d(X_n, X) > \varepsilon] \to 0$ as $n \to \infty$ where $I[\omega \in A]$ is the indicator function of the event $\omega \in A$. The statement $Y_n = O_P(1)$ means that the sequence Y_n is bounded in probability, i.e., for every $\varepsilon > 0$, there exist $M, N > 0$ such that $PI[d(X_n, 0) < M] > 1 - \varepsilon$ for $n > N$.

Understanding claims such as $o_P(1) + o_P(1) = o_P(1)$ or $O_P(1)o_P(1) = o_P(1)$ is easiest when each order symbol is replaced with an appropriate sequence. The first example is a statement of the continuous mapping theorem, i.e., $X_n + Y_n = Z_n \xrightarrow{\text{p}} 0$ where X_n, Y_n both converge to zero in probability. The second statement can be seen as a consequence of combining Prohorov's theorem and Slutsky's theorem, i.e., if X_n is bounded in probability and $Y_n \xrightarrow{\text{p}} 0$, then $X_n Y_n \xrightarrow{\text{p}} 0$.

We interpret $o_{\text{as}}(1)$ and $O_{\text{as}}(1)$ similar to the above, however the convergence is understood in the almost sure sense, i.e., for all $\varepsilon > 0$, $PI[\lim_{n\to\infty} d(X_n, X) < \varepsilon] = 1$. In the case of the almost sure order notation, the appropriate distribution shall be understood from the context.

2.3.2 Main Definitions & Theorems

Empirical process theory extends statements like the Strong Law of Large Numbers and the Central Limit Theorem to be uniform in a class of functions, $f \in \mathcal{F}$, and

serves as an elegant stepping stone for proving the convergence of m-estimators.

To establish the the uniform convergence of $\{P_n f : f \in \mathcal{F}\}$, i.e., $\sup_{f\in\mathcal{F}} \|P_n f - Pf\| \xrightarrow{\text{P}} 0$, it is necessary to characterize the size of the class $\mathcal{F}$. This notion can be made precise under the following definitions.

A measurable function $F : \mathcal{X} \to \mathbb{R}$ such that $|f(x)| \leq F(x)$ for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$ is called an *envelope function* for a class of functions $\mathcal{F}$. A function $f : \mathcal{X} \to \mathbb{R}$ is said to be μ-integrable, or simply *integrable*, when $\int |f(x)|d\mu(x) < \infty$. Denote the supremum norm of a function $\phi : \mathcal{F} \to \mathbb{R}$ by $\|\phi\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\phi(f)|$. We denote the $L_r(P)$-norm (of f) as $\|f\|_{P,r} = (P|f|^r)^{1/r}$.

Definition 1. Given two measurable functions $l, u : \mathcal{X} \to \mathbb{R}$ with $X \sim P$ and finite $L_r(P)$-norm, the bracket $[l, u]$ is the collection of all functions $f \in \mathcal{F}$ such that $l(x) \leq f(x) \leq u(x)$ for all $x \in \mathcal{X}$. An *ε-bracket* is a bracket such that $\|u - l\| < \varepsilon$ for some norm $\|\cdot\|$ (defined for collections).

Definition 2. The *bracketing number* $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the smallest number of ε-brackets needed to cover $\mathcal{F}$.

Intuitively, the bracketing number is the minimum number "intervals" needed to cover every co-domain of every function of a family.

Definition 3. A collection $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ with $\|P_n f - Pf\|_{\mathcal{F}} = o_{\text{as}}(1)$ is said to be P-Glivenko–Cantelli.

Theorem 1. *If $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|_{P,1}) < \infty$ for every $\varepsilon > 0$ then $\mathcal{F}$ is P-Glivenko–Cantelli.*

Proof. The proof is a straight-forward generalization of the Glivenko–Cantelli Theorem which states that the empirical distribution function converges almost surely. See [49] Theorem 19.1 for proof. □

Theorem 2. *If $\mathcal{F}$ is a Glivenko–Cantelli class of functions $f : \mathcal{X} \to \mathbb{R}^k$ with integrable envelope and $\phi : \mathbb{R}^k \to \mathbb{R}$ is continuous, then the class of functions $\phi \circ f : \mathcal{X} \to \mathbb{R}$ is Glivenko–Cantelli, provided this class has an integrable envelope.*

For proof of Theorem 2, see [50].

Although the proofs of Theorems 1 and 2 are omitted, they are straightforward consequences of the the Glivenko-Cantelli Theorem. This theorem however is certainly non-trivial; its understanding is not essential to this thesis. It is sufficient to understand that if the co-domains of a family are covered by finitely many, small intervals, then the average of any member of this family almost surely converges to its expectation.

2.4 Statistical Characterization of the SME

In this section we relate the stochastic m-estimator (5) to the population maximizer (7). Following classical statistical analysis, we make this characterization in the limit of large data, i.e., $n \to \infty$ where n is the number of samples.

Logically speaking, this this characterization is not indicative of the estimator's performance in any real-world setting. However, we may reasonably regard this asymptotic conclusion as being a desirable, common-sense property. Moreover, the $n \to \infty$ assumption vastly simplifies the analysis and allows an elegant characterization of the SME under conditions which are typically relatively easy to verify.

Presently we show that the SME–being based on a finite training set–would indeed recover the population maximizer as $n \to \infty$. This is the property of "statistical consistency," or simply "consistency." Usually the maximizer of (5) is found as the $\theta \in \Theta$ which is a zero of the gradient (of (5)). We also show that such a procedure results in an SME which is Normally distributed in the limit of n. As we see in later chapters, this conclusion allows us to precisely characterize the risk associated with different parameterizations of SMEs, i.e., the risk as a function of λ and/or β.

2.4.0.1 A General Consistency Proof

For the moment, we abstract away the details of (5) and establish the conditions of consistency for a "non-specific m-estimator." That is, we examine conditions of $\{m_\theta :$

$\theta \in \Theta\}$ which ensure that $\hat{\theta}_n = \arg\max_{\theta\in\Theta} P_n m_\theta(X)$ is almost surely arbitrarily close to a member of Θ_0 defined as $\Theta_0 = \{\theta \in \Theta : Pm_\theta(X) = \max_{\theta'\in\Theta} Pm_{\theta'}(X)\}$. We temporarily defer our analysis of the particular specification $m_\theta(X) = \int_{\mathbb{R}} Z(t)m_{\theta,t}(\theta)\,\mathrm{d}\nu(t)$.

Theorem 3. *Suppose that the class of functions $\{m_\theta : \theta \in \Theta\}$ is P-Glivenko–Cantelli and that there exists $\Theta_0 \subseteq \Theta$ such that $Pm_{\theta_0} = Pm_{\theta_0'}$ for all $\theta_0, \theta_0' \in \Theta_0$ and that $\sup\{Pm_\theta : d(\theta, \Theta_0) \geq \varepsilon\} < Pm_{\theta_0}$ for every $\varepsilon > 0$. Then $P_n m_{\hat{\theta}_n} \geq P_n m_{\theta_0}$ implies that $d(\hat{\theta}_n, \Theta_0) \xrightarrow{\text{as}} 0$.*

Proof. By the property of $\hat{\theta}_n$ and the strong law of large numbers, we have $P_n m_{\hat{\theta}_n} \geq P_n m_{\theta_0} = Pm_{\theta_0} - o_{\mathsf{as}}(1)$. Subtracting $Pm_{\hat{\theta}_n}$ from both sides and rearranging gives,

$$\begin{aligned} Pm_{\theta_0} - Pm_{\hat{\theta}_n} &\leq P_n m_{\hat{\theta}_n} - Pm_{\hat{\theta}_n} + o_{\mathsf{as}}(1) \\ &\leq \sup_{\theta\in\Theta} |P_n m_\theta - Pm_\theta| + o_{\mathsf{as}}(1) \\ &\xrightarrow{\text{as}} 0, \end{aligned}$$

where the almost sure convergence follows from the definition of Glivenko-Cantelli.

By assumption, there exists for every $\varepsilon > 0$ a number $\eta > 0$ such that $Pm_\theta < Pm_{\theta_0} - \eta$ for every θ with $d(\theta, \Theta_0) > \varepsilon$. Thus, the event $\{d(\hat{\theta}_n, \theta_0) \geq \varepsilon\}$ is contained in the event $\{Pm_{\hat{\theta}_n} < Pm_{\theta_0} - \eta\}$. The probability of the latter event converges to zero almost surely, in view of the preceding display. This proof can be found in [49]. □

Clearly the simplicity of this proof is a consequence of the "high-level" assumption that $\{m_\theta : \theta \in \Theta\}$ is a P-Glivenko–Cantelli class. We now turn to establishing this condition for the stochastic m-estimator.

2.4.0.2 SME is Glivenko–Cantelli

We begin by noting that (5) is simply an m-estimator evaluated over an augmentation of data X. That is, writing the augmented data as $Y = (X, Z)$ where X, Z have the

same meaning as above, we can rewrite (5) as

$$\theta \mapsto P_n m_\theta(Y)$$
$$m_\theta(Y) = \int_{\mathbb{R}} Z(t) m_{\theta,t}(X)\, \mathrm{d}\nu(t)$$

where P_n is again the empirical distribution of $\{Y^{(i)}\}_1^n$. The population criterion has the form, $\theta \mapsto P_X P_{Z|X,\lambda} m_\theta(Y)$.

We now connect Wald's classical regularity conditions to connect the SME (5) to Definition 3. Unlike Wald, we weaken the identifiability requirement and allow the global maximum to be achieved at (possibly) several points, i.e., $|\Theta_0| \geq 1$.

Lemma 1. *Let (Θ, d) be a compact metric space, let the map $\theta \mapsto m_\theta(x)$ be continuous for every $x \in \mathcal{X}$, and suppose that every θ has a neighborhood U such that $\sup_{\theta' \in U} |m_{\theta'}(x)| \leq K_\theta(x)$ where $\|K_\theta(x)\|_{\mu,1} < \infty$. Then the class $\{m_\theta : \theta \in \Theta\}$ is Glivenko–Cantelli and $\sup\{Pm_\theta : d(\theta, \Theta_0) \geq \varepsilon\} < Pm_{\theta_0}$ for every $\varepsilon > 0$ and any $\theta_0 \in \Theta_0$ if and only if $\theta \mapsto Pm_\theta$ attains its global maximum only on points Θ_0.*

Proof. The compactness of Θ and the local domination of the functions m_θ imply that the class $\{m_\theta : \theta \in \Theta\}$ possesses an integrable envelope function. The dominated convergence and the assumed continuity of the maps $\theta \mapsto m_\theta(x)$ imply that the map $\theta \mapsto Pm_\theta$ is continuous.

The set $B_\varepsilon = \{\theta \in \Theta : d(\theta, \Theta_0) \geq \delta\}$ is an intersection of closed and bounded sets (since Θ is bounded). Since $\theta \mapsto Pm_\theta$ is continuous and B_ε compact, $\theta \mapsto Pm_\theta$ attains its supremum in B_ε for every given $\varepsilon > 0$. By assumption, this maximum is smaller than Pm_{θ_0}.

To complete the proof we show that,

$$N_{[]}(\varepsilon, \{m_\theta : \theta \in \Theta\}, \|\cdot\|_{P,1}) < \infty,$$

and invoke Thm. 1. If B_m is a decreasing sequence of neighborhoods of a fixed θ such that $\cap_m B_m = \{\theta\}$ and u_m and l_m are defined as the supremum and infimum of the

functions m_θ with $\theta \in B_m$, then $u_m - l_m \to m_\theta - m_\theta = 0$ as $m \to \infty$, by the continuity of the functions $\theta \to m_\theta$, By the dominated convergence theorem $P(u_m - l_m) \to 0$. We conclude that for every $\varepsilon > 0$ and $\theta \in \Theta$ there exists a neighborhood B such that $P(u_B - l_B) < \varepsilon$ where $u_B = \sup_{\theta' \in B} m_{\theta'}$ and $l_B = \inf_{\theta' \in B} m_{\theta'}$. The collection of neighborhoods B obtained this way by varying θ over Θ has finite sub-collection that covers Θ, by the compactness of Θ. The corresponding finitely-many brackets $[l_B, u_B]$ cover the class $\{m_\theta : \theta \in \Theta\}$. This proof can be found in [50]. □

The integrable bound assumption of Lemma 1 is satisfied when $\sup_{\theta' \in U} \|z(t)m_{\theta',t}(x)\|_{\nu,1} \leq K(x,z)$ where $\|K(x,z)\|_{\mu,1} < \infty$. This is often the most difficult condition to verify. In this thesis we rarely need the generality afforded by (5) and in the subsequent chapters we typically ensure,

- z is bounded, i.e., $\sup_t |z(t)\mathrm{d}\nu(t)| = c < \infty$, and,
- z is a finite-length vector, i.e., $|\{t : \nu(t) > 0\}| < \infty$.

It then remains to prove that the component m-functions are such that, $\sup_{\theta' \in U} |m_{\theta',t}(x)| \leq K_t(x)$, since,

$$\begin{aligned}\sup_{\theta' \in U} \|z(t)m_{\theta',t}(x)\|_{\nu,1} &= \sup_{\theta' \in U} \sum_{t \in T} z_t m_{\theta',t}(x)\mathrm{d}\nu(t) \\ &\leq c|T| \max_{t \in T} \sup_{\theta' \in U} K_t(x),\end{aligned}$$

where $T = \{t : \nu(t) > 0\}$. These rather strong assumptions allow for the immediate disregard of the additional randomness of Z since we need only replace the finitely many $z(t)$ with the supremum. It is also useful note that when $\|Z\|_{\nu,1} < \infty$ and $\|\sup_{\theta \in U} m_{\theta,t}\|_{\nu(t),1} < \infty$, the Cauchy-Schwartz inequality ensures the existence of the integrable bound.

2.4.1 Asymptotic Normality

Having established that the SME $\hat{\theta}_n = \arg\max_{\theta\in\Theta} P_n m_\theta(X)$ is eventually near the statistically ideal set Θ_0, we now attempt to characterize the nature of this convergence. Specifically, we ask what is the distribution of the random estimator $\hat{\theta}_n$?

Since we wish to not preclude the possibility of multiple points of maximum $|\Theta_0| > 1$, we briefly introduce the Z-estimator as a framework to describe and examine the local convergence. The z-estimator is defined as the $\hat{\theta}_n$ which satisfies $\|P_n\psi_{\hat{\theta}_n}\| = 0$. That this estimator is closely related to the m-estimator is apparent by the fact that $\{\theta \in \Theta : \|P_n\psi_\theta(X)\| = 0\}$ includes all extrema of $P_n m_\theta(X)$ when $\psi_\theta(X) = \nabla_\theta m_\theta(X)$. Hence, the appropriate subset $K \subseteq \Theta$ implies equivalence between z- and m- estimators.

Like the m-estimator consistency proof, we employ empirical process theory as a framework for providing a simple proof of the z-estimator's. consistency

Theorem 4. *Suppose that the class of functions $\{\psi_\theta : \theta \in \Theta\}$ is P-Glivenko–Cantelli and that there exists $\emptyset \neq \Theta_0 \subseteq \Theta$ such that $\sup\{\|P\psi_\theta\| : d(\theta, \Theta_0) > \delta\} > 0 = \|P\psi_{\theta_0}\|$ for every $\delta > 0$. Then $P_n\psi_{\theta_{\hat{\theta}_n}} = 0$ implies that $d(\hat{\theta}_n, \Theta_0) \xrightarrow{\text{as}} 0$.*

Proof. By Theorem 2, $\|P\psi_{\hat{\theta}}\| = \|P_n\psi_{\hat{\theta}}\| + o(1) = o(1)$, almost surely as $n \to \infty$, by the proper of $\hat{\theta}$. Thus it is impossible that $d(\hat{\theta}, \theta_0) > \delta$ infinitely often, for any $\delta > 0$. □

We now prove the asymptotic Normality of the z-estimator, and by appropriate restriction of Θ, the m-estimator as well.

Theorem 5. *For each θ in an open subset of Euclidean space, let $\theta \mapsto \psi_\theta(x)$ be twice continuously differentiable for every x. Suppose that $P\psi_{\theta_0} = 0$, that $P||\psi_{\theta_0}||^2 < \infty$ and that the matrix $P\dot{\psi}_{\theta_0}$ exists and is non-singular. Assume that the second-order partial derivatives are dominated by a fixed integrable function $\ddot{\psi}(x)$ for every θ in a*

neighborhood of θ_0*. Then every consistent estimator sequence* $\hat{\theta}_n$ *such that* $\Psi_n(\hat{\theta}_n) = 0$ *for every* n *satisfies*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -(P\dot{\psi}_{\theta_0})^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\theta_0}(X_i) + o_P(1).$$

In particular, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *is asymptotically normal with mean 0 and covariance matrix* $(P\dot{\psi}_{\theta_0})^{-1}P\psi_{\theta_0}\psi_{\theta_0}^{\mathsf{T}}(P\dot{\psi}_{\theta_0})^{-1}$.

Proof. By Taylor's theorem there exists a (random) vector $\tilde{\theta}_n$ on the line segment between θ_0 and $\hat{\theta}_n$ such that

$$0 = \Phi_n(\hat{\theta}_n) = \Phi_n(\theta_0) + \dot{\Phi}_n(\theta_0)(\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^{\mathsf{T}}\ddot{\Psi}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0).$$

The first term on the right $\Psi_n(\theta_0)$ is an average of the iid random vectors $\psi_{\theta_0}(X_i)$, which have mean $P\Psi_{\theta_0} = 0$, By the central limit theorem, the sequence $\sqrt{n}\Phi_n(\theta_0)$ converges in distribution to a multivariate normal distribution with mean 0 and covariance matrix $P\psi_{\theta_0}\psi_{\theta_0}^{\mathsf{T}}$. The derivative $\dot{\Phi}_n(\theta_0)$ in the second term is an average also. By the law of large numbers it converges in probability to the matrix $V = P\dot{\psi}_\theta$. The second derivative $\ddot{\Phi}_n(\tilde{\theta}_n)$ is a k-vector of $(k \times k)$ matrices depending on the second-order derivatives $\ddot{\psi}_n$. By assumption, there exists a ball B around θ_0 such that $\ddot{\psi}_\theta$ is dominated by $||\ddot{\psi}||$ for every $\theta \in \mathsf{B}$. The probability of the event $\{\hat{\theta}_n \in \mathsf{B}\}$ tends to 1. On this event

$$\|\ddot{\Psi}_n(\tilde{\theta}_0)\| = \|\frac{1}{n}\sum_{i=1}^{n}\ddot{\psi}_{\tilde{\theta}_n}(X_i)\| \leq \frac{1}{n}\sum_{i=1}^{n}\|\ddot{\psi}(X_i)\|.$$

This is bounded in probability by the law of large numbers. Combination of these facts allows us to rewrite the preceding display as

$$-\Psi_n(\theta_0) = (V + o_P(1) + \frac{1}{2}(\hat{\theta}_n - \theta_0)O_P(1))(\hat{\theta}_n - \theta_0) = (V + o_P(1))(\hat{\theta}_n - \theta_0),$$

because the sequence $(\hat{\theta}_n - \theta_0)O_P(1) = o_P(1)O_P(1)$ converges to 0 in probability if $\hat{\theta}_n$ is consistent for θ_0. The probability that the matrix $V_{\theta_0} + o_P(1)$ is invertible tends to

1. Multiply the preceding equation by $\sqrt{n}$ and apply $(V + o_P(1))^{-1}$ on the left and right to complete the proof. This proof is due to Wald and presented in [49]. $\square$

Assuming then that interchange of differentiation and integration is reasonable, we see that the SME is asymptotically normal with covariance $\Upsilon^{-1}\Sigma\Upsilon^{-1}$, where,

$$\Upsilon = P \int_{\mathbb{R}} Z(t)\ddot{m}_{\theta,t}(X)\,\mathrm{d}\nu(t)$$
$$\Sigma = P\left(\int_{\mathbb{R}} Z(t)\dot{m}_{\theta,t}(X)\,\mathrm{d}\nu(t)\right)\left(\int_{\mathbb{R}} Z(t)\dot{m}_{\theta,t}(X)\,\mathrm{d}\nu(t)\right)^{\mathsf{T}}.$$

CHAPTER III

COMPUTATIONAL COST: LEARNING OBSERVED MRFS

In this chapter we develop a particular stochastic m-estimator which is best suited for undirected, high tree-width graphical models. Such graphical models contain loops (where parameter inference is concerned); often there are many edges with respect to the number of random variables. We will assume that the data is fully observed, that is, the data generating distribution is naturally regarded as not being a marginal of some other distribution. As we will demonstrate, the application of this estimator to high tree-width graphs is a matter of practical concern. Low tree-width graphs usually permit computationally efficient algorithms for exact inference and need not be approximated.

3.1 Introduction

Maximum likelihood estimation is by far the most popular point estimation technique in machine learning and statistics. Assuming that the data consists of n, m-dimensional vectors

$$D = (X^{(1)}, \ldots, X^{(n)}), \quad X^{(i)} \in \mathbb{R}^m, \tag{8}$$

and is sampled iid from a parametric distribution p_{θ_0} with $\theta_0 \in \Theta \subset \mathbb{R}^r$, a maximum likelihood estimator (MLE) $\hat{\theta}_n^{\mathrm{ml}}$ is a maximizer of the log-likelihood function

$$\ell_n(\theta\,; D) = \sum_{i=1}^{n} \log p_\theta(X^{(i)}) \tag{9}$$

$$\hat{\theta}_n^{\mathrm{ml}} = \arg\max_{\theta \in \Theta} \ell_n(\theta\,; D). \tag{10}$$

As an m-estimator, the use of the MLE is motivated by its consistency,[1] i.e., $\hat{\theta}_n^{\text{ml}} \to \theta_0$ as $n \to \infty$ with probability 1 [24]. The consistency property ensures that as the number n of samples grows, the estimator will converge to the true parameter θ_0 governing the data generation process.

An even stronger motivation for the use of the MLE is that it has an asymptotically normal distribution with mean vector θ_0 and variance matrix $(nI(\theta_0))^{-1}$. More formally, we have the following convergence in distribution as $n \to \infty$ [24]

$$\sqrt{n}\,(\hat{\theta}_n^{\text{ml}} - \theta_0) \rightsquigarrow N(0, I^{-1}(\theta_0)), \tag{11}$$

where $I(\theta)$ is the $r \times r$ Fisher information matrix

$$I(\theta) = \mathsf{E}_{p_\theta}\{\nabla \log p_\theta(X)(\nabla \log p_\theta(X))^\top\} \tag{12}$$

with ∇f representing the $r \times 1$ gradient vector of $f(\theta)$ with respect to θ. The convergence (11) is especially striking since according to the Cramer-Rao lower bound, the asymptotic variance $(nI(\theta_0))^{-1}$ of the MLE is the smallest possible variance for any estimator. Since it achieves the lowest possible asymptotic variance, the MLE (and other estimators which share this property) is said to be asymptotically efficient.

The consistency and asymptotic efficiency of the MLE motivate its use in many circumstances. Unfortunately, in some situations the maximization or even evaluation of the log-likelihood (9) and its derivatives is impossible due to computational considerations. For instance this is the situation in many high dimensional exponential family distributions, including Markov random fields whose graphical structure contains cycles. This has lead to the proposal of alternative estimators under the premise that a loss of asymptotic efficiency is acceptable–in return for reduced computational complexity.

[1]The consistency $\hat{\theta}_n^{\text{ml}} \to \theta_0$ with probability 1 is sometimes called strong consistency in order to differentiate it from the weaker notion of consistency in probability $P(|\hat{\theta}_n^{\text{ml}} - \theta_0| < \epsilon) \to 0$.

In contrast to asymptotic efficiency, we view consistency as a less negotiable property and prefer to avoid inconsistent estimators if at all possible. This common viewpoint in statistics is somewhat at odds with recent advances in the machine learning literature promoting non-consistent estimators, for example using variational techniques [31]. Nevertheless, we feel that there is a consensus regarding the benefits of having consistent estimators over non-consistent ones [53, 54].

In this chapter, we propose a family of estimators for use in situations where the computation of the MLE is intractable. In contrast to many previously proposed approximate estimators, these estimators are statistically consistent and admit a precise quantification of both computational complexity and statistical accuracy through their asymptotic variance. Due to the continuous parameterization of the estimator family, we obtain an effective framework for achieving a predefined problem-specific balance between computational tractability and statistical accuracy. We also demonstrate that in some cases reduced computational complexity may in fact act as a regularizer, increasing robustness and therefore accomplishing both reduced computation and increased accuracy. This "win-win" situation conflicts with the conventional wisdom stating that moving from the MLE to pseudo likelihood and other related estimators result in a computational win but a statistical loss [32]. Nevertheless we show that this occurs in some practical situations.

For the sake of concreteness, we focus on the case of estimating the parameters associated with Markov random fields. In this case, we provide a detailed discussion of the accuracy–complexity tradeoff. We include experiments on both simulated and real world data for several models including the Boltzmann machine, conditional random fields, and the Boltzmann linear chain model.

3.2 Related Work

There is a large body of work dedicated to tractable learning techniques. Two popular categories are Markov chain Monte Carlo (MCMC) and variational methods. MCMC is a general purpose technique for approximating expectations and can be used to approximate the normalization term and other intractable portions of the log-likelihood and its gradient [13]. Variational methods are techniques for conducting inference and learning based on tractable bounds [31]. A similar approach would be to conduct maximum likelihood estimation for a simpler model that is tractable. Variational methods are most useful when the inference distribution is a marginal of some joint distribution which is an exponential family [54].

Despite the substantial work on MCMC and variational methods, there are little practical results concerning the convergence and approximation rate of the resulting parameter estimators. Variational techniques are sometimes inconsistent and it is hard to analyze their asymptotic statistical behavior [55]. In the case of MCMC, a number of asymptotic results exist [13], but since MCMC plays a role inside each gradient descent or EM iteration it is hard to analyze the asymptotic behavior of the resulting parameter estimates. An advantage of this framework is that we are able to directly characterize the asymptotic behavior of the estimator and relate it to the amount of computational savings.

This work draws on the composite likelihood method for parameter estimation proposed by [36] which in turn generalized the pseudo likelihood of [6]. A selection of more recent studies on pseudo and composite likelihood are [1, 34, 51, 48, 28]. Most of the recent studies in this area examine the behavior of the pseudo or composite likelihood in a particular modeling situation. We believe that the present work is the first to systematically examine statistical and computational tradeoffs in a general quantitative framework. Possible exceptions are [60] which is an experimental study on texture generation, [58] which is focused on inference rather than parameter

Table 1: High-level comparison of the dominant techniques for learning MRFs. Here consistency and accuracy are understood in the statistical sense, hence accuracy is dominated by asymptotic variance. It is arguable to what extent the methods facilitate a computation/accuracy tradeoff, but certainly only SCL manages a smooth tradeoff.

	Consistent	Accuracy*	Computation*	Tradeoff	Smooth
Variational Methods	✗	✗	✓	✓	✗
Markov Chain Monte Carlo	✓	✗	✓	✓	✗
Contrastive Divergence	✗	✗	✓	✓	✗
Full Likelihood	✓	✓	✓	✗	✗
Pseudo Likelihood	✓	✓	✓	✗	✗
Composite Likelihood	✓	✓	✓	✓	✗
Stochastic Composite Likelihood	✓	✓	✓	✓	✓

$*$ - "Feasible to analyze/valuate" not a statement of being.

estimation, and [10] which examines the generalization performance of small- and large- scale learning systems. The work of [35] is also interesting in that the authors employ composite likelihood m-estimators and asymptotic arguments to compare the risk of discriminative and generative models. However, this work differs in theme and technique–we explore the tradeoff between computation and accuracy by way of a fundamentally different estimator.

Composite likelihood techniques, and consequently this work, can be thought of as local contrastive objectives (i.e., pseudo likelihood, contrastive divergence). [52] present a non-local alternative in which the objective is not restricted to using the training label, but rather any assignment.

3.3 Stochastic Composite Likelihood

In many cases, the absence of a closed form expression for the normalization term prevents the computation of the log-likelihood (9) and its derivatives thereby severely limiting the use of the MLE. A popular example is Markov random fields, wherein the

computation of the normalization term is often intractable (see Section 3.6 for more details). In this work we propose alternative estimators based on the maximization of a stochastic variation of the composite likelihood.

We denote multiple samples using superscripts and individual dimensions using subscripts. Thus $X_j^{(r)}$ refers to the j-dimension of the r sample. Following standard convention we refer to random variables (RV) using uppercase letters and their corresponding values using lowercase letters. We also use the standard notations for extracting a subset of the dimensions of a random variable

$$X_S \stackrel{\text{def}}{=} \{X_i : i \in S\}, \qquad X_{-j} \stackrel{\text{def}}{=} \{X_i : i \neq j\}. \tag{13}$$

We start by reviewing the pseudo log-likelihood function [6] associated with the data D (8),

$$p\ell_n(\theta\,; D) \stackrel{\text{def}}{=} \sum_{i=1}^{n}\sum_{j=1}^{m} \log p_\theta(X_j^{(i)}|X_{-j}^{(i)}). \tag{14}$$

The maximum pseudo likelihood estimator (MPLE) $\hat{\theta}_n^{\text{mpl}}$ is consistent, i.e., $\hat{\theta}_n^{\text{mpl}} \to \theta_0$ with probability 1, but possesses considerably higher asymptotic variance than the MLE's $(nI(\theta_0))^{-1}$ [39]. Its main advantage is that it does not require the computation of the normalization term as it cancels out in the probability ratio defining conditional distributions

$$p_\theta(X_j|X_{-j}) = p_\theta(X_j|\{X_k : k \neq j\}) = \frac{p_\theta(X)}{\sum_{x_j} p_\theta(X_1, \ldots, X_{j-1}, X_j = x_j, X_{j+1}, \ldots, X_m)}. \tag{15}$$

The MLE and MPLE represent two different ways of resolving the tradeoff between asymptotic variance and computational complexity. The MLE has low asymptotic variance but high computational complexity while the MPLE has higher asymptotic variance but low computational complexity. It is desirable to obtain additional estimators realizing alternative resolutions of the accuracy complexity tradeoff. To this end we define the stochastic composite likelihood whose maximization provides a family of consistent estimators with statistical accuracy and computational complexity spanning the entire accuracy-complexity spectrum.

Stochastic composite likelihood generalizes the likelihood and pseudo likelihood functions by constructing an objective function that is a stochastic sum of likelihood objects. We start by defining the notion of m-pairs and likelihood objects and then proceed to stochastic composite likelihood.

Definition 4. An m-pair (A, B) is a pair of sets $A, B \subset \{1, \ldots, m\}$ satisfying $A \neq \emptyset = A \cap B$. The likelihood object associated with an m-pair (A, B) and X is $S_\theta(A, B) \stackrel{\text{def}}{=} \log p_\theta(X_A|X_B)$ where X_S is defined in (13). The composite log-likelihood function [36] is a collection of likelihood objects defined by a finite sequence of m-pairs $(A_1, B_1), \ldots, (A_k, B_k)$

$$c\ell_n(\theta\,; D) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \sum_{j=1}^{k} \log p_\theta(X^{(i)}_{A_j}|X^{(i)}_{B_j}). \tag{16}$$

There is a certain lack of flexibility associated with the composite likelihood framework as each likelihood object is either selected or not for the entire sample $X^{(1)}, \ldots, X^{(n)}$. There is no allowance for some objects to be selected more frequently than others. For example, available computational resources may allow the computation of the log-likelihood for 20% of the samples, and the pseudo likelihood for the remaining 80%. In the case of composite likelihood if we select the full likelihood component (or the pseudo likelihood or any other likelihood object) then this component is applied to all samples indiscriminately.

In SCL, different likelihood objects $S_\theta(A_j, B_j)$ may be selected for different samples with the possibility of some likelihood objects being selected for only a small fraction of the data samples. The selection may be non-coordinated, in which case each component is selected or not independently of the other components. Or it may be coordinated in which case the selection of one component depends on the selection of the other ones. For example, we may wish to avoid selecting a pseudo likelihood component for a certain sample $X^{(i)}$ if the full likelihood component was already selected for it.

Another important advantage of stochastic selection is that the discrete parameterization of (16) defined by the sequence $(A_1, B_1), \ldots, (A_k, B_k)$ is less convenient for theoretical analysis. Each component is either selected or not, turning the problem of optimally selecting components into a hard combinatorial problem. The stochastic composite likelihood, which is defined below, enjoys continuous parameterization leading to more convenient optimization techniques and convergence analysis.

Definition 5. Consider a finite sequence of m-pairs $(A_1, B_1), \ldots, (A_k, B_k)$, a dataset $D = (X^{(1)}, \ldots, X^{(n)})$, $\beta \in \mathbb{R}^k_+$, and n iid, length k, binary random vectors $\{Z^{(i)}\}_{i=1}^n \overset{\text{iid}}{\sim} P(Z)$ with $\lambda_j \overset{\text{def}}{=} \mathsf{E}(Z_j) > 0$. The stochastic composite log-likelihood (SCL) is

$$sc\ell_n(\theta\,; D, Z) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} m_\theta(X^{(i)}, Z^{(i)}), \quad \text{where} \tag{17}$$

$$m_\theta(X, Z) \overset{\text{def}}{=} \sum_{j=1}^{k} \beta_j Z_j \log p_\theta(X_{A_j} | X_{B_j}), \tag{18}$$

where, for brevity, we typically omit D, Z in favor of $sc\ell_n(\theta)$.

In other words, the SCL is a stochastic extension of (16) where for each sample $X^{(i)}, i = 1, \ldots, n$, the likelihood objects $S(A_1, B_1), \ldots, S(A_k, B_k)$ are either selected or not, depending on the values of the binary random variables $Z_1^{(i)}, \ldots, Z_k^{(i)}$ and weighted by the constants $\beta_1, \ldots, \beta_m$. Note that $Z_j^{(i)}$ may in general depend on $Z_r^{(i)}$ but not on $Z_r^{(l)}$ or on $X^{(i)}$.

When we focus on examining different models for $P(Z)$ we sometimes parameterize it, for example by λ, i.e., $P_\lambda(Z)$. This reuse of λ (it is also used in Definition 5) is a notational abuse. We accept it, however, as in most of the cases that we consider $\lambda_1, \ldots, \lambda_k$ from Definition 5 either form the parameter vector for $P(Z)$ or are part of it. Often we refer to a particular λ as a "policy" in order to emphasize its role as a "knob" in selecting particular m-pairs.

Some illustrative examples follow.

Independence. Factorizing $P_\lambda(Z_1, \ldots, Z_k) = \prod_j P_{\lambda_j}(Z_j)$ leads to $Z_j^{(i)} \sim \text{Ber}(\lambda_j)$ with complete independence among the indicator variables. For each sample $X^{(i)}$, each likelihood object $S(A_j, B_j)$ is selected or not independently with probability λ_j.

Multinomial. A multinomial model $Z \sim \text{Mult}(1, \lambda)$ implies that for each sample $Z^{(i)}$ a multivariate Bernoulli experiment is conducted with precisely one likelihood object being selected depending on the selection probabilities $\lambda_1, \ldots, \lambda_k$.

Product of Multinomials. A product of multinomials is formed by a partition of the dimensions to l disjoint subsets $\{1, \ldots, m\} = C_1 \cup \cdots C_l$ where $Z_{C_i} \sim \text{Mult}(1, (\lambda_j : j \in C_i))$, i.e.,

$$P(Z) = \prod_{i=1}^{c} P_i\left(\{Z_j : j \in C_i\}\right), \qquad \text{where } P_i \text{ is } \text{Mult}(1, (\lambda_j : j \in C_l)).$$

Loglinear Models. The distribution $P(Z)$ follows a hierarchical loglinear model [7]. This case subsumes the other cases above.

In analogy to the MLE and the MPLE, the maximum SCL estimator (MSCLE) $\hat{\theta}_n^{\text{msl}}$ estimates θ_0 by maximizing the SCL function. In contrast to the log-likelihood and pseudo log-likelihood functions, the SCL function and its maximizer are random variables that depend on the indicator variables $Z^{(1)}, \ldots, Z^{(n)}$ in addition to the data D. As such, its behavior should be summarized by examining the limit $n \to \infty$. Doing so eliminates the dependency on particular realizations of $Z^{(1)}, \ldots, Z^{(n)}$ in favor of the the expected frequencies $\lambda_j = \mathsf{E}_{P(Z)} Z_j$ which are non-random constants.

The statistical accuracy and computational complexity of the MSCL estimator are continuous functions of the parameters (β, λ) (components weights and selection probabilities respectively) which vary continuously throughout their domain $(\lambda, \beta) \in \Lambda \times \mathbb{R}_+^k$. Choosing appropriate values of (λ, β) retrieves the special cases of MLE, MPLE, maximum composite likelihood with each selection being associated with

a distinct statistical accuracy and computational complexity. The SCL framework allows selections of many more values of (λ, β) realizing a wide continuous spectrum of estimators, each resolving the accuracy-complexity tradeoff differently.

We include below a demonstration of the SCL framework in a simple low dimensional case. In the following sections we discuss in detail the statistical behavior of the MSCLE and its computational complexity. We conclude the chapter with several experimental studies.

3.3.1 Boltzmann Machine Example

Before proceeding we illustrate the SCL framework using a simple example involving a Boltzmann machine [27]. Section 3.8.1 describes the specifics of this model. We consider in detail three SCL policies: full likelihood (FL), pseudo likelihood (PL), and a stochastic combination of first and second order pseudo likelihood with the first order components $(p(X_i|X_{-i}))$ selected with probability λ and the second order components $(p(X_i, X_j|X_{\{i,j\}^c}))$ with probability $1-\lambda$.

Denoting the number of (binary) graph vertices, or nodes, by m, the number of examples by n, the computational complexity of the FL function (FLOP[2] counts) is (for the log-likelihood) and $O\big(\binom{m}{2}(2^m+n)\big)$ (for the log-likelihood) and $O\big(\binom{m}{2}^2 2^m + n\binom{m}{2}\big)$ (for the log-likelihood gradient).[3] The exponential growth in m prevents such computations for large graphs.

The k-order PL function offers a practical alternative to FL (1-order PL correspond to the traditional pseudo likelihood and 2-order is its analog with second order components $p(X_{\{i,j\}}|X_{\{i,j\}^c})$). The complexity of computing the corresponding SCL function is $O\big(\binom{m}{2}(\binom{m}{k}2^k+n)\big)$ (for the objective function) and $O\big(\binom{m}{2}^2\binom{m}{k}2^k + n\binom{m}{2}\big)$ (for the gradient). The slower complexity growth of the k-order PL (polynomial in m

[2]Number of FLoating point OPerations.

[3]With memoization the complexity of the gradient can be reduced to $O\big(\binom{m}{2}2^m + n\binom{m}{2}\big)$ (at the cost of exponential 2^m storage). Note that this is only a polynomial improvement to an exponential complexity hence we lose no insight by making naive assumptions.

instead of exponential) is offset by its reduced statistical accuracy, which we measure using the normalized asymptotic variance

$$\text{eff}(\hat{\theta}_n) = \frac{\det(\text{Asymp Var}(\hat{\theta}_n))}{\det(\text{Asymp Var}(\hat{\theta}_n^{\text{ml}}))} \tag{19}$$

which is bounded from below by 1 (due to Cramer Rao lower bound) and its deviation from 1 reflects its inefficiency relative to the MLE.

The MLE thus achieves the best accuracy but it is computationally intractable. The first order and second order PL have higher asymptotic variance but are easier to compute. The SCL framework enables adding many more estimators filling in the gaps between ML, 1-order PL, 2-order PL, etc.

We illustrate three SCL functions in the context of a simple Boltzmann machine (five binary nodes, fourteen samples $X^{(1)}, \ldots, X^{(14)}$, $\theta^{\text{true}} = (-1, -1, -1, -1, -1, 1, 1, 1, 1, 1)$) in Figure 2. The top box refers to the full likelihood policy, i.e., maximum likelihood. For each of the fourteen samples, the FL component is computed and their aggregation forms the SCL function which in this case equals the log-likelihood. The selection of the FL component for each sample is illustrated using a diamond box. The numbers under the boxes reflect the FLOP counts needed to compute the components and the total complexity associated with computing the entire SCL or log-likelihood is listed on the right. As mentioned above, the normalized asymptotic variance (19) is 1.

The pseudo likelihood function (14) is illustrated in the second box where each row correspond to one of the five PL components. As each of the five PL component is selected for each of the samples we have diamond boxes covering the entire 5×14 array. The shade of the diamond boxes reflects the complexity required to compute them enabling an easy comparison to the FL components in the top of the figure (note how the FL boxes are much darker than the PL boxes). The numbers at the bottom of each column reflect the FLOP marginal count for each of the fourteen samples and the numbers to the right of the rows reflect the FLOP marginal count for each of the

PL components. In this case the FLOP count is less than half the FLOP count of the FL in top box (this reduction in complexity obtained by replacing FL with PL will increase dramatically for graphs with more than 5 nodes) but the asymptotic variance is 83% higher.[4]

The third SCL policy reflects a stochastic combination of first and second order pseudo likelihood components. Each first order component is selected with probability λ and each second order component is selected with probability $1 - \lambda$. The result is a collection of 5 1-order PL components and 10 2-order components with only some of them selected for each of the fourteen samples. Again diamond boxes correspond to selected components which are shaded according to their FLOP complexity. The per-component FLOP marginals and per example FLOP marginals are listed as the bottom row and right-most column. The total complexity is somewhere between FL and PL and the asymptotic variance is reduced from the PL's 183% to 148%.

Additional insight may be gained at this point by considering Figure 4 which plots several SCL estimators as points in the plane whose x and y coordinates correspond to normalized asymptotic variance and computational complexity respectively. We turn at this point to considering the statistical properties of the SCL estimators.

3.4 Consistency and Asymptotic Variance of $\hat{\theta}_n^{msl}$

A nice property of the SCL framework is enabling mathematical characterization of the statistical properties of the estimator $\hat{\theta}_n^{\text{msl}}$. In this section we examine the conditions for consistency of the MSCLE and its asymptotic distribution and in the next section we consider robustness. The propositions below constitute novel generalizations of some well-known results in classical statistics. Proofs may be found in Appendix A. For simplicity, we assume that X is discrete and $p_\theta(x) > 0$.

Although we could merely appeal to the statistical guarantees outlined in Chapter

[4]The asymptotic variance of SCL functions is computed using formulas derived in the next section.

		$X^{(1)}$	$X^{(2)}$	$X^{(3)}$	$X^{(4)}$	$X^{(5)}$	$X^{(6)}$	$X^{(7)}$	$X^{(8)}$	$X^{(9)}$	$X^{(10)}$	$X^{(11)}$	$X^{(12)}$	$X^{(13)}$	$X^{(14)}$	Total
FL	$X_1, , \ldots, X_5$	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	4620
	Complexity	330	330	330	330	330	330	330	330	330	330	330	330	330	330	**4620**
	Rel.Efficiency	**1.00**														
PL	$X_1 \vert X_{-1}$	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	308
	$X_2 \vert X_{-2}$	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	308
	$X_3 \vert X_{-3}$	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	308
	$X_4 \vert X_{-4}$	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	308
	$X_5 \vert X_{-5}$	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	◇	308
	Complexity	110	110	110	110	110	110	110	110	110	110	110	110	110	110	**1540**
	Rel.Efficiency	**1.83**														
0.7PL+0.3PL2	$X_1 \vert X_{-1}$		◇	◇	◇	◇				◇	◇	◇	◇			176
	$X_2 \vert X_{-2}$		◇		◇	◇		◇	◇	◇	◇	◇		◇	◇	220
	$X_3 \vert X_{-3}$	◇	◇				◇	◇	◇		◇	◇	◇	◇	◇	220
	$X_4 \vert X_{-4}$			◇			◇	◇				◇	◇	◇	◇	154
	$X_5 \vert X_{-5}$	◇				◇	◇	◇	◇	◇		◇	◇		◇	198
	$X_{\{1,2\}} \vert X_{\{1,2\}^c}$			◇	◇		◇					◇				164
	$X_{\{1,3\}} \vert X_{\{1,3\}^c}$	◇		◇		◇				◇			◇			205
	$X_{\{1,4\}} \vert X_{\{1,4\}^c}$					◇	◇	◇			◇					164
	$X_{\{1,5\}} \vert X_{\{1,5\}^c}$	◇						◇				◇	◇			164
	$X_{\{2,3\}} \vert X_{\{2,3\}^c}$				◇	◇		◇			◇	◇				205
	$X_{\{2,4\}} \vert X_{\{2,4\}^c}$		◇	◇				◇	◇		◇	◇	◇			287
	$X_{\{2,5\}} \vert X_{\{2,5\}^c}$	◇			◇		◇		◇							164
	$X_{\{3,4\}} \vert X_{\{3,4\}^c}$	◇							◇							82
	$X_{\{3,5\}} \vert X_{\{3,5\}^c}$			◇		◇	◇		◇							164
	$X_{\{4,5\}} \vert X_{\{4,5\}^c}$							◇	◇	◇	◇		◇			205
	Complexity	208	107	208	167	230	230	293	271	148	230	274	252	66	88	**2772**
	Rel.Efficiency	**1.48**														

Figure 2: Sample runs of three different SCL policies for 14 examples $X^{(1)}, \ldots, X^{(14)}$ drawn from a 5 binary node Boltzmann machine ($\theta^{\text{true}} = (-1, -1, -1, -1, -1, 1, 1, 1, 1, 1)$). The policies are full likelihood (FL, top), pseudo likelihood (PL, middle), and a stochastic combination of first and second order pseudo likelihood with the first order components selected with probability 0.7 and the second order components with probability 0.3 (bottom).

The sample runs for the policies are illustrated by placing a diamond box in table entries corresponding to selected likelihood objects (rows corresponding to likelihood objects and columns to $X^{(1)}, \ldots, X^{(14)}$). The FLOP counts of each likelihood object determines the shade of the diamond boxes while the total FLOP counts per example and per likelihood objects are displayed as table marginals (bottom row and right column for each policy). We also display the total FLOP count and the normalized asymptotic variance (19).

Even in the simple case of 5 nodes, FL is the most complex policy with PL requiring a third of the FL computation. 0.7PL+0.3PL2 is somewhere in between. The situation is reversed for the estimation accuracy-FL achieves the lowest possible normalized asymptotic variance of 1, PL is almost twice that, and 0.7PL+0.3PL2 somewhere in the middle. The SCL framework spans the accuracy-complexity spectrum. Choosing the right λ value obtains an estimator that suits available computational resources and required accuracy.

2, we choose to develop proofs which exploit the specific nature of composite likelihood m-functions. This has two advantages. The proofs are self-contained and easier to understand as we do not need generality afforded by empirical process theory. Second, we wish to establish a stronger conclusion, viz., that the SCL estimator converges to the MLE under the appropriate conditions.

Definition 6. A sequence of m-pairs $(A_1, B_1), \ldots, (A_k, B_k)$ is m-pair identifiable, or simply identifiable, of p_θ if the map $\{p_\theta(X_{A_j}|X_{B_j}) : j = 1, \ldots, k\} \mapsto p_\theta(X)$ is injective. In other words, there exists only a single collection of conditionals $\{p_\theta(X_{A_j}|X_{B_j}) : j = 1, \ldots, k\}$ that does not contradict the joint $p_\theta(X)$.

Proposition 1. *Let $\Theta \subset \mathbb{R}^r$ be an open set, $p_\theta(x) > 0$ and continuous and smooth in θ, and $(A_1, B_1), \ldots, (A_k, B_k)$ be a sequence of m-pairs for which $\{(A_j, B_j) : \forall j$ such that $\lambda_j > 0\}$ ensures identifiability. Then the sequence of SCL maximizers is strongly consistent, i.e.,*

$$P\left(\lim_{n\to\infty} \hat{\theta}_n = \theta_0\right) = 1. \tag{20}$$

The above proposition indicates that to guarantee consistency, the sequence of m-pairs needs to satisfy Definition 6. It can be shown that a selection equivalent to the pseudo likelihood function, i.e.,

$$\mathcal{S} = \{(A_1, B_1), \ldots, (A_m, B_m)\} \quad \text{where} \quad A_i = \{i\}, B_i = \{1, \ldots, m\} \setminus A_i \tag{21}$$

ensures identifiability and consequently the consistency of the MSCLE estimator. Furthermore, every selection of m-pairs that subsumes $\mathcal{S}$ in (21) similarly guarantees identifiability and consistency.

The proposition below establishes the asymptotic normality of the MSCLE $\hat{\theta}_n$. The asymptotic variance enables the comparison of SCL functions with different parameterizations (λ, β).

Proposition 2. *Making the assumptions of Proposition 1 as well as convexity of* $\Theta \subset \mathbb{R}^r$ *we have the following convergence in distribution*

$$\sqrt{n}(\hat{\theta}_n^{msl} - \theta_0) \rightsquigarrow N\left(0, \Upsilon\Sigma\Upsilon\right) \tag{22}$$

where,

$$\Upsilon^{-1} = \sum_{j=1}^{k} \beta_j \lambda_j \, \mathsf{Var}_{\theta_0}(\nabla S_{\theta_0}(A_j, B_j)) \tag{23}$$

$$\Sigma = \mathsf{Var}_{\theta_0}\left(\sum_{j=1}^{k} \beta_j \lambda_j \nabla S_{\theta_0}(A_j, B_j)\right). \tag{24}$$

The notation $\mathsf{Var}_{\theta_0}(Y)$ represents the covariance matrix of the random vector Y under p_{θ_0} while the notations $\xrightarrow{\text{p}}, \rightsquigarrow$ in the proof below denote convergences in probability and in distribution [24]. ∇ represents the gradient vector with respect to θ.

When θ is a vector the asymptotic variance is a matrix. To facilitate comparison between different estimators we follow the convention of using the determinant, and in some cases the trace, to measure the statistical accuracy. See chapter 4 of [44] for some heuristic arguments for doing so. Figures 2,3,4 provide the asymptotic variance for some SCL estimators and describe how it can be used to gain insight into which estimator to use.

The fact that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to a Gaussian with zero mean (for the MLE and similarly for SCL estimators as we show above) implies that the estimator's asymptotic behavior, up to $n^{-1/2}$ order, is determined exclusively by the asymptotic variance. That means that the estimator is essentially unbiased up to that order. Higher order statistical analysis (obtained using Taylor series with more terms) show that the bias decays in the faster rate of n^{-1} [18]. We thus follow the statistical convention of conducting first order asymptotic analysis and concentrate on the estimator's asymptotic variance.

The statistical accuracy of the SCL estimator depends on β (weight parameters) and λ (selection parameter). It is thus desirable to use the results in this section in determining what values of β, λ to use. Directly using the asymptotic variance is not possible in practice as it depends on the unknown quantity θ_0. However, it is possible to estimate the asymptotic variance using the training data. We describe this in Section 3.7.

3.4.1 Assessing Risk

So far we have demonstrated that estimators based on maximizing the likelihood, conditional composite likelihood, and stochastic composite likelihood criterion functions can be interpreted as stochastic m-estimators. Accordingly, when the model is well-specified, i.e., $p \equiv p_{\theta_0}$ for a unique $\theta_0 \in \Theta$, all three estimators recover the true model in the limit of large data. Alternatively, when the model is misspecified, i.e., there exists no $\theta_0 \in \Theta$ such that $p \equiv p_{\theta_0}$, then the MLE recovers a different model than either the SCL or CL objectives.

Table 2 summarizes the relationship between maximum likelihood (MLE), composite likelihood (MCLE), and stochastic composite likelihood estimators (MSCLE). In the well-specified case, i.e., $p \equiv p_{\theta_0}$ for some $\theta_0 \in \Theta$, all three estimators share the same point of convergence. In the misspecified case, i.e., there exists no $\theta_0 \in \Theta$ for which $p \equiv p_{\theta_0}$ then the MLE recovers a possibly different point in the limit of large data. That the MCLE and MSCLE recover the same point is matter of appropriately chosen weights, i.e., setting the SCL weights to $\beta_j = 1/\lambda_j$ adjusts for sampling bias with respect to the unweighted composite likelihood.

Despite this precise characterization of each estimator, it remains unclear which estimator performs better in practice. A partial answer to this question was given by the estimator's asymptotic distribution. In each of these cases, the estimator is asymptotically Normal, hence we heuristically characterized the estimator based on

Table 2: Limit points of estimators based on maximizing likelihood, conditional composite likelihood, and stochastic composite likelihood criterion functions. Well-specified corresponds to existence of some $\theta \in \Theta$ such that $p \equiv p_{\theta_0}$ while misspecified indicates no such $\theta \in \Theta$ exists.

	Well-specified	Misspecified
MLE	$\hat{\theta}_n^{\mathsf{ml}} \xrightarrow{\text{as}} \theta_0$	$\hat{\theta}_n^{\mathsf{ml}} \xrightarrow{\text{as}} \theta_0^{\mathsf{ml}}$
MCLE	$\hat{\theta}_n^{\mathsf{cl}} \xrightarrow{\text{as}} \theta_0$	$\hat{\theta}_n^{\mathsf{cl}} \xrightarrow{\text{as}} \theta_0^{\mathsf{cl}}$
MSCLE	$\hat{\theta}_n^{\mathsf{scl}} \xrightarrow{\text{as}} \theta_0$	$\hat{\theta}_n^{\mathsf{scl}} \xrightarrow{\text{as}} \theta_0^{\mathsf{cl}}$

some scalar-valued function of its covariance matrix, e.g., $\mathsf{tr}\,\Sigma$ or $\det \Sigma$.

We now explore these choices by examining the risk associated with each estimator. Herein, it is assumed that the testing task entails inference over a subset of the random vector, i.e., infer Y based on some evidence X.

In the standard case of the MLE, risk is often characterized in terms of expected conditional log-loss, i.e.,

$$\mathcal{R}(\theta) = \mathsf{E}_{p(X,Y)} \log p_\theta(Y|X).$$

However, for many MRF's

We write the total error of an estimator, as

$$\mathcal{R}(\hat{\theta}_n) - \mathcal{R}_p = \Big(\mathcal{R}(\hat{\theta}_n) - \mathcal{R}(\theta_0)\Big) + \Big(\mathcal{R}(\theta_0) - \mathcal{R}_p\Big), \tag{25}$$

where $\mathcal{R}_p = H[p(Y|X)]$, or Bayes risk, is the entropy of the true conditional distribution. It is important to emphasize that the point θ_0 is not necessarily the point with minimum expected log-loss, but rather is the maximizer of some objective taken under the population distribution, i.e., $\theta_0 = \arg\max_{\theta \in \Theta} M(\theta)$. The parenthetical decomposition of risk reflects the understanding that some of the risk is due to estimation error, i.e., error due to making inferences from a finite set of samples, and approximation error, i.e., error arising from the objective's inability to adequately represent nature.

3.4.1.1 Main Result

Proposition 3. *Let Σ^{-1} be the asymptotic variance as defined in Prop. 2. Assume $\mathcal{R}(\theta)$ is a real-valued twice continuously differentiable function. Write $\mathcal{R}_n = \mathcal{R}(\hat{\theta}_n)$ and $\mathcal{R}_0 = \mathcal{R}(\theta_0)$. Then, under the assumptions of Prop. 2.*

$$\sqrt{n}(\mathcal{R}_n - \mathcal{R}_0) \rightsquigarrow \mathsf{N}\left(0, \dot{\mathcal{R}}_0^T \Sigma^{-1} \dot{\mathcal{R}}_0\right). \tag{26}$$

Furthermore, if $\dot{\mathcal{R}}_0 = 0$ then,

$$n(\mathcal{R}_n - \mathcal{R}_0) \rightsquigarrow \sum_{i=1}^{d} w_i \chi_1^2 \tag{27}$$

where the right-hand is the distribution associated with a weighted sum of χ^2 random variables with $w_i = [A^{\frac{1}{2}} \Sigma^{-1} A^{\frac{1}{2}}]_{ii} - [B^{\frac{1}{2}} \Sigma^{-1} B^{\frac{1}{2}}]_{ii}$, for some positive-definite A and positive semi-definite B such that $\ddot{\mathcal{R}}_0 = A - B$.

Proof. The proof uses a standard argument known as the delta method and follows [49, 35].

We use the stochastic order notion $o_p(\cdot)$ to indicate that for a sequence of random variables R_n, the statement $X_n = o_p(R_n)$ means $X_n = Y_n R_n$ for $Y_n \xrightarrow{\text{p}} 0$.

For the first statement, we take the first-order Taylor expansion of $\dot{\mathcal{R}}_n$ on the neighborhood of θ_0,

$$\mathcal{R}_n = \mathcal{R}_0 + \dot{\mathcal{R}}_0^{\mathsf{T}}(\hat{\theta}_n - \theta_0) + o_p(\|\hat{\theta}_n - \theta_0\|). \tag{28}$$

That this is valid, follows from Taylor's theorem and the fact that the continuous mapping theorem implies that if $R(h) = o(\|h\|^r)$ for $h \to 0$ and $r > 0$, then $R(W_n) = o_p(\|W_n\|^r)$ for any $W_n = o_p(1)$. In this case, the consistency of $\hat{\theta}_n$ implies $\hat{\theta}_n - \theta_0 = o_p(1)$.

Since $\sqrt{n}(\hat{\theta}_n - \theta_0)$ was found to be asymptotically Normal, we may invoke Prohorov's theorem and conclude that the sequence is uniformly tight, from which it follows that, $o_p(\sqrt{n}\|\hat{\theta}_n - \theta_0\|) = o_p(1)$.

Multiplying (28) by $\sqrt{n}$ and rearranging terms yields,

$$\sqrt{n}(\mathcal{R}_n - \mathcal{R}_0) = \dot{\mathcal{R}}_0\sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(1).$$

Equation (26) follows from application of Slutsky's theorem.

When $\dot{\mathcal{R}}_0 = 0$, Equation 26 is a degenerate distribution.

Similarly to the previous procedure, we take the second-order Taylor expansion of $\dot{\mathcal{R}}_0$ near θ_0, multiply by n, and rearrange terms to get,

$$\begin{aligned} n(\mathcal{R}_n - \mathcal{R}_0) &= \tfrac{1}{2}\sqrt{n}(\hat{\theta}_n - \theta_0)^{\mathsf{T}}\ddot{\mathcal{R}}\sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(1) \\ &= \tfrac{1}{2}\,\mathsf{tr}\left\{\sqrt{n}(\hat{\theta}_n - \theta_0)^{\mathsf{T}}(A - B)\sqrt{n}(\hat{\theta}_n - \theta_0)\right\} + o_p(1) \\ &= \tfrac{1}{2}\,\mathsf{tr}\left([A^{\frac{1}{2}}\sqrt{n}(\hat{\theta}_n - \theta_0)]^{\otimes}\right) \\ &\quad - \tfrac{1}{2}\,\mathsf{tr}\left([B^{\frac{1}{2}}\sqrt{n}(\hat{\theta}_n - \theta_0)]^{\otimes}\right) + o_p(1), \end{aligned}$$

where the $v^{\otimes} = vv^{\mathsf{T}}$ denotes an outer product of v.

Since $\ddot{\mathcal{R}}_0^{\frac{1}{2}}\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N(0, \ddot{\mathcal{R}}_0^{\frac{1}{2}}\Sigma\ddot{\mathcal{R}}_0^{\frac{1}{2}})$, application of the continuous mapping theorem to the outer product function implies a Wishart limiting distribution, hence,

$$\begin{aligned} n(\mathcal{R}_n - \mathcal{R}_0) \rightsquigarrow \tfrac{1}{2}\,\mathsf{tr}\,\mathsf{W}\left(A^{\frac{1}{2}}\Sigma^{-1}A^{\frac{\mathsf{T}}{2}}, 1\right) - \\ \tfrac{1}{2}\,\mathsf{tr}\,\mathsf{W}\left(B^{\frac{1}{2}}\Sigma^{-1}B^{\frac{\mathsf{T}}{2}}, 1\right), \end{aligned} \tag{29}$$

where $\mathsf{W}(V, n)$ is the Wishart distribution with n degrees of freedom.

Finally we note that, $\mathsf{tr}\,\mathsf{W}(V, 1) = \mathsf{tr}\,V\mathsf{W}(I, 1)$, which is the distribution of a weighted sum of independent χ_1^2 variables, where the weights are determined by the diagonal elements of V. Hence (29) follows. □

As discussed in [35], (29) can be understood in the following sense. Let $V = \ddot{\mathcal{R}}_0^{\frac{1}{2}}\Sigma^{-1}\ddot{\mathcal{R}}_0^{\frac{\mathsf{T}}{2}}$. The mean of this distribution is $\frac{1}{2}\,\mathsf{tr}\,(V)$ and the variance is $\mathsf{tr}\,(V * V)$, where $*$ is the element-wise product.

3.5 Robustness of $\hat{\theta}_n^{msl}$

The experimental results of the SCL estimator exhibited the surprising phenomenon of sometimes out-performing the maximum likelihood estimator on a held-out test set (see Section 3.8). This phenomenon seems to be in contradiction to the fact that the asymptotic variance of the MLE is lower than that of the SCL maximizer. This is explained by the fact that in some cases the true model generating the data does not lie within the parametric family $\{p_\theta : \theta \in \Theta\}$ under consideration. For example, many graphical models (HMM, CRF, LDA, etc.) make conditional independence assumptions that are often violated in practice. In such cases the SCL estimator acts as a regularizer achieving better test set performance than the non-regularized MLE. We provide below a theoretical account of this phenomenon by considering the statistical development of Chapter 2.

As Chapter 2 was a theoretical presentation of the abstract SME and lacked detailed examples, we re-establish that chapter's earlier proofs in the specific context of the composite likelihood SME. Our notation here continues to follow the one in [49].

We now assume that the model generating the data is outside the model family $P(X) \notin \{p_\theta : \theta \in \Theta\}$ and we extend the notation of $m_\theta(X, Z)$ in (18) with,

$$\begin{aligned}
\psi_\theta(X, Z) &\stackrel{\text{def}}{=} \nabla m_\theta(X, Z) \\
\dot{\psi}_\theta(X, Z) &\stackrel{\text{def}}{=} \nabla^2 m_\theta(X, Z) \quad \text{(matrix of second order derivatives)} \\
\Psi_n(\theta) &\stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \psi_\theta(X^{(i)}, Z^{(i)}).
\end{aligned}$$

Proposition 4 below generalizes the consistency result by asserting that $\hat{\theta}_n \to \theta_0$ where θ_0 is the point on $\{p_\theta : \theta \in \Theta\}$ that is closest to the true model P, as defined

by

$$\theta_0 = \arg\max_{\theta\in\Theta} M(\theta) \quad \text{where} \quad M(\theta) \stackrel{\text{def}}{=} -\sum_{j=1}^{k} \beta_j \lambda_j D(P(X_{A_j}|X_{B_j})||p_\theta(X_{A_j}|X_{B_j})), \tag{30}$$

or equivalently, θ_0 satisfies

$$\mathsf{E}_{P(X)}\, \mathsf{E}_{P(Z)}\, \psi_{\theta_0}(X, Z) = 0. \tag{31}$$

When the SCL function reverts to the log-likelihood function, θ_0 becomes the KL projection of the true model P onto the parametric family $\{p_\theta : \theta \in \Theta\}$.

Proposition 4. *Assuming the conditions in Proposition 1 as well as* $\sup_{\theta:\|\theta-\theta_0\|\geq\epsilon} M(\theta) < M(\theta_0)$ *for all* $\epsilon > 0$ *we have* $\hat{\theta}_n^{msl} \to \theta_0$ *as* $n \to \infty$ *with probability 1.*

The added condition maintains that θ_0 is a well separated maximum point of M. In other words it asserts that only values close to θ_0 may yield a value of M that is close to the maximum $M(\theta_0)$. This condition is satisfied in the case of most exponential family models.

Proposition 5. *Assuming the conditions of Proposition 2 as well as* $\mathsf{E}_{P(X)}\, \mathsf{E}_{P(Z)}\, \|\psi_{\theta_0}(X, Z)\|^2 < \infty$, $\mathsf{E}_{P(X)}\, \mathsf{E}_{P(Z)}\, \dot{\psi}_{\theta_0}(X)$ *exists and is non-singular,* $|\ddot{\Psi}_{ij}| = |\partial^2\psi_\theta(x)/\partial\theta_i\theta_j| < g(x)$ *for all* i, j *and* θ *in a neighborhood of* θ_0 *for some integrable* g*, we have*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -(\mathsf{E}_{P(X)}\, \mathsf{E}_{P(Z)}\, \dot{\psi}_{\theta_0})^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\theta_0}(X^{(i)}, Z^{(i)}) + o_P(1) \tag{32}$$

or equivalently

$$\hat{\theta}_n = \theta_0 - (\mathsf{E}_{P(X)}\, \mathsf{E}_{P(Z)}\, \dot{\psi}_{\theta_0})^{-1} \frac{1}{n} \sum_{i=1}^{n} \psi_{\theta_0}(X^{(i)}, Z^{(i)}) + o_P\left(\frac{1}{\sqrt{n}}\right). \tag{33}$$

Above, $f_n = o_P(g_n)$ means f_n/g_n converges to 0 with probability 1.

Corollary 1. *Assuming the conditions specified in Proposition 5 we have*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N(0, (\mathsf{E}_{P(X)}\, \mathsf{E}_{P(Z)}\, \dot{\psi}_{\theta_0})^{-1}(\mathsf{E}_{P(X)}\, \mathsf{E}_{P(Z)}\, \psi_{\theta_0}\psi_{\theta_0}^\top)(\mathsf{E}_{P(X)}\, \mathsf{E}_{P(Z)}\, \dot{\psi}_{\theta_0})^{-1}). \tag{34}$$

Equation (33) means that asymptotically, $\hat{\theta}_n$ behaves as θ_0 plus the average of iid RVs. As mentioned in [49] this fact may be used to obtain a convenient expression for the asymptotic influence function, which measures the effect of adding a new observation to an existing large dataset. Neglecting the remainder in (32) we have

$$\begin{aligned}
\mathcal{I}(x,z) &\stackrel{\text{def}}{=} \hat{\theta}_n(X^{(1)},\ldots,X^{(n-1)},x,Z^{(1)},\ldots,Z^{(n-1)},z) - \hat{\theta}_{n-1}(X^{(1)},\ldots,X^{(n-1)},Z^{(1)},\ldots,Z^{(n-1)}) \\
&\approx -(\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\dot{\psi}_{\theta_0})^{-1}\left(\frac{1}{n}\sum_{i=1}^{n-1}\psi_{\theta_0}(X^{(i)},Z^{(i)}) + \frac{1}{n}\psi_{\theta_0}(w,z) - \frac{1}{n-1}\sum_{i=1}^{n-1}\psi_{\theta_0}(X^{(i)},Z^{(i)})\right) \\
&= -(\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\dot{\psi}_{\theta_0})^{-1}\frac{1}{n}\psi_{\theta_0}(w,z) + (\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\dot{\psi}_{\theta_0})^{-1}\frac{1}{n(n-1)}\sum_{i=1}^{n-1}\psi_{\theta_0}(X^{(i)},Z^{(i)}) \\
&= -\frac{1}{n}(\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\dot{\psi}_{\theta_0})^{-1}\psi_{\theta_0}(w,z) + o_P\left(\frac{1}{n}\right). \qquad (35)
\end{aligned}$$

Corollary 1 and Equation 35 measure the statistical behavior of the estimator when the true distribution is outside the model family. In these cases it is possible that a computationally efficient SCL maximizer will result in higher statistical accuracy as well. This "win-win" situation of improving in both accuracy and complexity over the MLE is confirmed by our experiments in Section 3.8.

We finally note that the above analysis is not limited to misspecified models. For example, the influence function may be used to detect the robustness of $\hat{\theta}_n$ to outliers or rare events (it is desirable to be robust to such occurrences even if the model is not misspecified).

3.6 Stochastic Composite Likelihood for Markov Random Fields

Markov random fields (MRF) are some of the more popular statistical models for complex high dimensional data. Approaches based on pseudo likelihood and composite likelihood are naturally well-suited in this case due to the cancellation of the normalization term in the probability ratios defining conditional distributions. More specifically, a MRF with respect to a graph $G = (V, E)$, $V = \{1,\ldots,m\}$ with a clique

set $\mathcal{C}$ is given by the following exponential family model

$$P_\theta(x) = \exp\left(\sum_{C\in\mathcal{C}} \theta_C f_C(x_C) - \log Z(\theta)\right),$$
$$Z(\theta) = \sum_x \exp\left(\sum_{C\in\mathcal{C}} \theta_c f_C(x_C)\right). \tag{36}$$

The primary bottlenecks in obtaining the maximum likelihood are the computations $\log Z(\theta)$ and $\nabla \log Z(\theta)$. Their computational complexity is exponential in the graph's treewidth and for many cyclic graphs, such as the Ising model or the Boltzmann machine, it is exponential in $|V| = m$.

In contrast, the conditional distributions that form the composite likelihood of (36) are given by (note the cancellation of $Z(\theta)$)

$$P_\theta(x_A|x_B) = \frac{\sum_{x'_{(A\cup B)^c}} \exp\left(\sum_{C\in\mathcal{C}} \theta_C f_C((x_A, x_B, x'_{(A\cup B)^c})_C)\right)}{\sum_{x'_{(A\cup B)^c}} \sum_{x''_A} \exp\left(\sum_{C\in\mathcal{C}} \theta_C f_C((x''_A, x_B, x'_{(A\cup B)^c})_C)\right)}. \tag{37}$$

whose computation is substantially faster. Specifically, the computation of (37) depends on the size of the sets A and $(A\cup B)^c$ and their intersections with the cliques in $\mathcal{C}$. In general, selecting small $|A_j|$ and $B_j = (A_j)^c$ leads to efficient computation of the composite likelihood and its gradient. For example, in the case of $|A_j| = l, |B_j| = m-l$ with $l \ll m$ we have that $k \leq m!/(l!(m-l)!)$ and the complexity of computing the $c\ell(\theta)$ function and its gradient may be shown to require time that is at most exponential in l and polynomial in m.

3.6.1 Ensuring Identifiability

Under fairly mild assumptions the stochastic composite likelihood estimator provides an asymptotically consistent method for recovering a specific limit point $\theta_0^{\mathsf{scl}} \in \Theta$. This follows from the fact that the MSCLE is a particular SME. However, consistency does not ensure that the estimate is "good," merely that its reliability increases with the quantity of training data.

This section addresses this issue by examining precisely how particular choices of components ensure that the limit point of SCLE is the same limit point as the MLE. In particular, we make precise the relationship from joint to conditionals and, critically, how the conditional distributions relate to the joint distribution. In particular we examine simple sufficient conditions which ensure a collection of components uniquely characterize a joint (when it exists).

3.6.1.1 From Joint to Conditional

Introductory treatments of conditional distributions define the conditional probability as the ratio of joint and marginal densities, i.e., $P(A|B) = P(A, B)/P(B)$ for sets $A, B \in \mathcal{A}$ where $(\mathcal{X}, \mathcal{A})$ is a measurable space and P a probability measure. The division by $P(B)$ ensures that $P(\cdot|B)$ is a probability measure and ensures zero measure is allocated to points outside B, i.e., $P(B^c|B) = 0$.

Although intuitive, this definition obviously requires that either $P(B) > 0$ or some limiting argument can be made which justifies the existence of $P(A|B)$ for $P(B_n) \to 0$ as n increases. That such reasoning is potentially error-prone is best underscored by the Borel–Kolmogorov paradox. The (apparent) paradox is related to selecting a point at random from the surface of the earth and how conditioning on seemingly equivalent events yields different conditional distributions. In particular, a point lying on the equator is uniformly distributed with respect to longitude while a point on the meridian is not (with respect to latitude). The paradox is that both seem to be referring to the same great circle, just under a different coordinate system.

The resolution of the Borel-Kolmogorov paradox lies in the inadmissibility of conditioning on an event of measure zero (be it equator or meridian). To correctly understand the nature of each conditional, one must view the meridian and equator as a limit of some continuous function, hence the conditional must be also seen as a limit.

The following theorem (Theorem 5.3.1 of [3]) makes the notion of conditional distribution precise by treating a it as a random variable, measurable under the law of the joint. In doing so, it shows that measurable conditional distributions always exist and that they are unique, almost everywhere.

Theorem 6. *Let $X : (\mathcal{X}, \mathcal{A}) \to (\mathcal{X}', \mathcal{A}')$ be a random object on $(\mathcal{X}, \mathcal{A}, P)$ and let B be a fixed set in $\mathcal{A}$. Then there exists a real-valued Borel measurable function $x \mapsto g(B|x)$ on $(\mathcal{X}', \mathcal{A}')$ such that for each $A \in \mathcal{A}'$,*

$$P(\{X \in A\} \cap B) = \int_A g(B|x) \mathrm{d}P(x). \tag{38}$$

Furthermore, if $x \mapsto h(B|x)$ is another such function, then $g \equiv h$, P-a.e.

Proof. The map $A \mapsto P(\{X \in A\} \cap B)$ is a finite measure on $\mathcal{A}'$ which is absolutely continuous with respect to P, hence the result follows from the Radon-Nikodym theorem. □

In simple terms, this theorem characterizes uniqueness and existence of the conditional distribution by viewing it as a random variable which partitions a measurable space. Then, through what is essentially the total law of probability, we see that the distribution must exist uniquely, P-almost everywhere. The uniqueness up to measure-zero provides some understanding of why the Borel-Kolmogorov conditionals are not necessarily in agreement. The paradox (and theorem) also reveals that conditional distributions are not invariant to transformations of the random variable– this would entail integration under a measure corresponding to the change of variable.

3.6.1.2 From Full Conditionals to Joint

We now consider conditions which ensure the converse of Section 3.6.1.1, viz., when does the map $\{(A_j, B_j)\}_1^k \mapsto \cup_j \{p(X_{A_j}|X_{B_j})\}$ characterize a joint distribution and is this distribution is unique. As outlined in Section 3.3 outlines, such conditions

have significant practical implications since conditional distributions of Markov random fields are typically computationally attractive while the corresponding joint is impossible to compute.

Brook's Lemma gives this converse, although under a fairly restrictive assumption known as the *positivity condition*, which we now state. [26, 11]

Definition 7. Let $(X_1, X_2, \ldots, X_m) \sim p(x_1, x_2, \ldots, x_m)$ and denote the marginal distribution of X_i as p_i. If $p_i(x_i) > 0$ for every $i = 1, \ldots, m$ implies that $p(x_1, \ldots, x_m) > 0$, then p is said to satisfy the *positivity condition*.

The positivity condition asserts that the support of the joint p is equivalent to the Cartesian product of the supports of the marginals, p_i's. That this condition should eliminate ambiguities like the Borel-Kolmogorov paradox follows from the fact that when the marginals have zero probability (meaning conditionals could be non-unique) then it follows that the joint must also have zero probability.

We also establish the term "compatible" to indicate that a pair of distributions agree on at least one joint distribution. More precisely,

Definition 8. Two conditional probability (density or mass) functions $p(X_{A_1}|X_{B_1}), q(X_{A_2}|X_{B_2})$ are said to be *compatible* if there exists marginals $\pi_1(X_{B_1}), \pi_2(X_{B_1})$ such that,

$$p(X_{A_1}|X_{B_1})\pi_1(X_{B_1}) \equiv q(X_{A_2}|X_{B_2})\pi_2(X_{B_2}).$$

[2]

We now illustrate how the positivity condition and simple factorization allows a joint to be written as a function of compatible conditionals.

Lemma 2 (Brook's Lemma)**.** *Let $k+1 : m-1$ denote the sequence $(k+1, k+2, \ldots, m-1)$. Under the positivity condition, a joint probability function p satisfies,*

$$p(x_1, \ldots, x_m) \propto \prod_{j=1}^{m} \frac{p(x_{\sigma(j)}|x_{\sigma(1:j-1)}, x_{\sigma(j+1:m)})}{p(x'_{\sigma(j)}|x_{\sigma(1:j-1)}, x'_{\sigma(j+1:m)})} \tag{39}$$

for every permutation $\sigma = (\sigma_1, \ldots, \sigma_m)$ *of* $\{1, \ldots, m\}$ *and any* $x' \in \mathcal{X}$ *(fixed for all* $x \in \mathcal{X}$*).*

Proof. Fix $x' \in \mathcal{X}$ and assume $\sigma = (1, 2, \ldots, m)$ is the identity permutation. The positivity assumption permits the factorization,

$$\begin{aligned} p(x_{1:m}) &= p(x_m|x_{1:m-1})f(x_{1:m-1}) \\ &= p(x_m|x_{1:m-1})\frac{p(x_{1:m-1}, x'_m)}{p(x'_m|x_{1:m-1})} \\ &= \frac{p(x_m|x_{1:m-1})}{p(x'_m|x_{1:m-1})}p(x_{1:m-1}, x'_m). \end{aligned}$$

Recursively making the expansion to right-hand joint shows that,

$$p(x_{1:m}) = p(x'_{1:m})\prod_{j=1}^{m}\frac{p(x_j|x_{1:j-1}, x'_{1:j-1})}{p(x'_j|x_{1:j-1}, x'_{1:j-1})}. \tag{40}$$

The argument is identical for an arbitrary permutation σ. □

It is important to emphasize that Brook's Lemma does not guarantee that the joint exists. As an example, consider two random variables each of which is drawn from a conditionally exponential distribution.

Example 1. Suppose that X_1, X_2 are random variables which, when viewed conditionally, have exponential distributions, i.e.,

$$\begin{aligned} X_1|x_2 &\sim \mathsf{Exp}(\lambda x_2) \\ X_2|x_1 &\sim \mathsf{Exp}(\lambda x_1). \end{aligned}$$

Lemma 2 implies,

$$\begin{aligned} f(x_1, x_2) &\propto \frac{f(x_1|x'_2)f(x_2|x_1)}{f(x'_1|x'_2)f(x'_2|x_1)} \\ &= \frac{\lambda x'_2 \exp(-\lambda x'_2 x_1)\lambda x_1 \exp(-\lambda x_1 x_2)}{\lambda x'_2 \exp(-\lambda x'_2 x'_1)\lambda x_1 \exp(-\lambda x_1 x'_2)} \\ &\propto \exp(-\lambda x_1 x_2). \end{aligned}$$

However, since the normalization term $\int_{\mathbb{R}^2_+} \exp(\lambda x_1 x_2)\mathrm{d}x$ is unbounded, no joint distribution exists. □

It is possible to extend Lemma 2 to distributions that fail to satisfy the positivity condition but additional assumptions are required.

Proposition 6. *Let f be a distribution defined on a finite state-space $|\mathcal{X}| < \infty$ of dimension m. Furthermore, assume that for every $(x, x') \in \mathcal{X}^2$ there exists an integer $m < \infty$ such that adjacent members of the sequence $(x^{(i)})_{i=1}^{m}$ differ only in a single component, i.e., there exists $j \leq m$ such that $x_j^{(i)} = x_j^{(i+1)}$, and that $f(x_j) > 0$. Then for a fixed $x' \in \mathcal{X}$, the factorization in Brook's Lemma holds.*

Proof. The proof consists of showing that the conditions imply ergodicity hence uniqueness. See [43] for proof. □

3.7 Automatic Selection of β

As Proposition 2 indicates, the weight vector β and selection probabilities λ play an important role in the statistical accuracy of the estimator through its asymptotic variance. The computational complexity, on the other hand, is determined by λ independently of β. Conceptually, we are interested in resolving the accuracy-complexity tradeoff jointly for both β, λ before estimating θ by maximizing the SCL function. However, since the computational complexity depends only on λ we propose the following simplified problem: Select λ based on available computational resources, and then given λ, select the β (and θ) that will achieve optimal statistical accuracy.

Selecting β that minimizes the asymptotic variance is somewhat ambiguous as $\Upsilon\Sigma\Upsilon$ in Proposition 2 is an $r \times r$ positive semidefinite matrix. A common solution is to consider the determinant as a one dimensional measure of the size of the variance matrix,[5] and minimize

$$J(\beta) = \log\det(\Upsilon\Sigma\Upsilon) = \log\det\Sigma + 2\log\det\Upsilon. \tag{41}$$

[5]See chapter 4 of [44] for a heuristic discussion motivating this measure.

A major complication with selecting β based on the optimization of (41) is that it depends on the true parameter value θ_0 which is not known at training time. This may be resolved, however, by noting that (41) is composed of covariance matrices under θ_0 which may be estimated using empirical covariances over the training set. To facilitate fast computation of the optimal β we also propose to replace the determinant in (41) with the product of the diagonal elements. Such an approximation is motivated by Hadamard's inequality (which states that for symmetric matrices $\det(M) \leq \prod_i M_{ii}$) and by Geršgorin's circle theorem (see below). This approximation works well in practice as we observe in the experiments section. We also note that the procedure described below involves only simple statistics that may be computed on the fly and does not contribute significant additional computation (nor do they require significant memory).

More specifically, we denote $K^{(ij)} = \mathsf{Cov}_{\theta_0}(\nabla S_{\theta_0}(A_i, B_i), \nabla S_{\theta_0}(A_j, B_j))$ with entries $K^{(ij)}_{st}$, and approximate the log det terms in (41) using

$$\log\det\Upsilon = -\log\det\sum_{j=1}^{k}\beta_j\lambda_j K^{(jj)} \approx -\sum_{l=1}^{r}\log\sum_{j=1}^{k}\beta_j\lambda_j K^{(jj)}_{ll} \tag{42}$$

$$\log\det\Sigma = \log\det\mathsf{Var}_{\theta_0}\left(\sum_{j=1}^{k}\beta_j\lambda_j\nabla S_{\theta_0}(A_j, B_j)\right) = \log\det\sum_{i=1}^{k}\sum_{j=1}^{k}\beta_i\lambda_i\beta_j\lambda_j K^{(ij)}$$
$$\approx \sum_{l=1}^{r}\log\sum_{i=1}^{k}\sum_{j=1}^{k}\beta_i\lambda_i\beta_j\lambda_j K^{(ij)}_{ll}. \tag{43}$$

We denote (assuming A is a $n \times n$ matrix) for $i \in \{1, \ldots, n\}$, $R_i(A) = \sum_{j\neq i}|A_{ij}|$ and let $D(A_{ii}, R_i(A))$ (D_i where unambiguous) be the closed disc centered at A_{ii} with radius $R_i(A)$. Such a disc is called a Geršgorin disc. The result below states that for matrices that are close to diagonal, the eigenvalues are close to the diagonal elements making our approximation accurate.

Theorem 3.7.1 (Geršgorin's circle theorem e.g., [29])**.** Every eigenvalue of A lies within at least one of the Geršgorin discs $D(A_{ii}, R_i(A))$. Furthermore, if the union of

k discs is disjoint from the union of the remaining $n-k$ discs, then the former union contains exactly k and the latter $n-k$ eigenvalues of A.

Algorithm 1 solves for θ, β jointly using alternating optimization. The second optimization problem $J(\beta;\cdot)$ is done using the approximation above and may be computed with minimal additional computation. The components of this objective are typically freely available when $sc\ell$ is minimized with Newton-type methods. In practice we found that such an approach leads to a selection of β that is close to optimal, despite loose convergence criteria for the minimization of the $sc\ell$ objective (see Sec. 3.8.3 and Figures 15, 21 for results).

Algorithm 1 Calculate $\hat{\theta}^{msl}$

Require: $\{X_i\}_{i\in I}$ and λ, $\beta^{(0)}$

$t \leftarrow 1$
while $t <$ MAXITS **do**
  $\theta^{(t)} \leftarrow \arg\min sc\ell(\theta; \{X_i\}_{i\in I}, \lambda, \beta^{(t-1)})$
  if converged **then**
    return θ
  end if
  $\beta^{(t)} \leftarrow \arg\min J(\beta; \{K^{(ij)}\}_{(i,j)\in J}, \lambda, \theta)$
  $t \leftarrow t+1$
end while
return false

3.8 Experiments

We demonstrate the asymptotic properties of $\hat{\theta}_n^{\text{msl}}$ and explore the complexity-accuracy tradeoff for three different models-Boltzmann machine, linear Boltzmann MRF and conditional random fields. In terms of datasets, we consider synthetic data as well as datasets from sentiment prediction and text chunking domains.

The basic road-map is to explore SCL for a theoretical Boltzmann machine and then to explore two datasets using both generative and discriminative models. We also demonstrate the effectiveness of the β heuristic for these experiments.

3.8.1 Toy Example: Boltzmann Machines

We illustrate the improvement in asymptotic variance of the MSCLE associated with adding higher order Boltzmann machine likelihood components with increasingly higher probability. The Boltzmann machine can be parameterized as,

$$p_\theta(x) = \exp\left(\sum_{i<j} \theta_{ij} x_i x_j - \log \psi(\theta)\right), \quad x \in \{0,1\}^m. \tag{44}$$

To be able to accurately compute the asymptotic variance we use $m = 5$ with θ being a $\binom{5}{2}$ dimensional vector with half the components $+1$ and half -1. Since the asymptotic variance of $\hat{\theta}_n^{\text{msl}}$ is a matrix we summarize its size using either its trace or determinant.

Figure 3 displays the asymptotic variance, relative to the minimal variance of the MLE, for the cases of full likelihood (FL), pseudo likelihood ($|A_j| = 1$) PL_1, stochastic combination of pseudo likelihood and 2nd order pseudo likelihood ($|A_j| = 2$) components $\alpha\text{PL}_2 + (1-\alpha)\text{PL}_1$, stochastic combination of 2nd order pseudo likelihood and 3rd order pseudo likelihood ($|A_j| = 3$) components $\alpha\text{PL}_3 + (1-\alpha)\text{PL}_2$, and stochastic combination of 3rd order pseudo likelihood and 4th order pseudo likelihood ($|A_j| = 4$) components $\alpha\text{PL}_4 + (1-\alpha)\text{PL}_3$.

The graph demonstrates the computation-accuracy tradeoff as follows: (a) pseudo likelihood is the fastest but also the least accurate, (b) full likelihood is the slowest but the most accurate, (c) adding higher order components reduces the asymptotic variance but also requires more computation, (d) the variance reduces with the increase in the selection probability α of the higher order component, and (e) adding 4th order components brings the variance very close the lower limit and with each successive improvement becoming smaller and smaller according to a law of diminishing returns.

Figure 4 displays the asymptotic accuracy and complexity for different SCL policies for $m = 9$ binary valued vertices of a Boltzmann machine. We explore three polices in which we denote pseudo likelihood components of size, or order, k. These

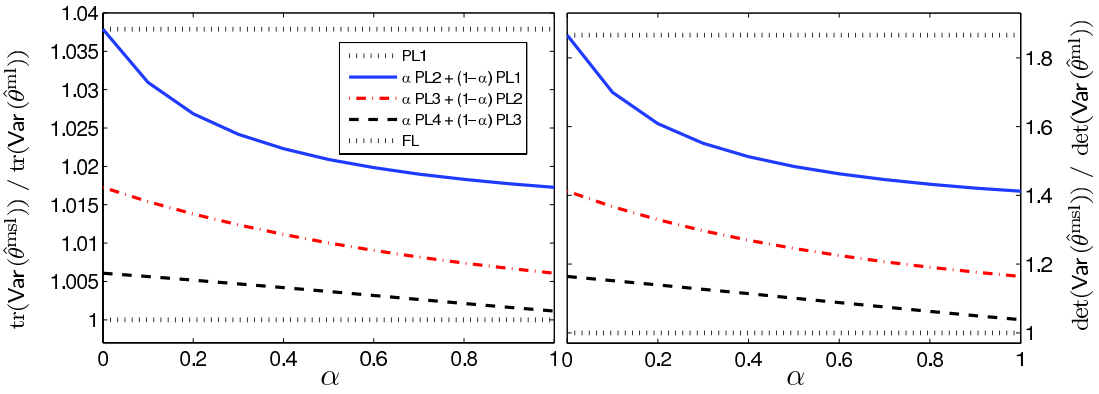

Figure 3: Asymptotic variance matrix, as measured by trace (left) and determinant (right), as a function of the selection probabilities for different stochastic versions of the SCL function.

policies include: $\lambda_1\beta_1\text{PL1} + \lambda_2(1-\beta_1)\text{PL2}$, $\lambda_1\beta_1\text{PL1} + \lambda_2(1-\beta_1)\text{PL3}$, $\lambda_1\beta_1\text{PL2} + \lambda_2(1-\beta_1)\text{PL3}$ (for multiple values of $\lambda_1, \lambda_2, \beta_1$). By taking different linear combinations of various sized pseudo likelihood components, we span a continuous spectrum of accuracy-complexity resolutions. The lower part of the diagram is the boundary of the achievable region (the optimal but unachievable place is the bottom left corner). SCL policies that lie to the right and top of that boundary may be improved by selecting a policy below and to the left of it.

3.8.2 Local Sentiment Prediction

Our first real world dataset experiment involves local sentiment prediction using a conditional MRF model. The dataset consisted of 249 movie review documents having an average of 30.5 sentences each with an average of 12.3 words from a 12633 word vocabulary. Each sentence was manually labeled as one of five sentimental designations: very negative, negative, objective, positive, or very positive. As described in [38] (where more information may be found) we considered the task of predicting the local sentiment flow within these documents using regularized conditional random fields (CRFs) (see Figure 5 for a graphical diagram of the model in the case of four sentences).

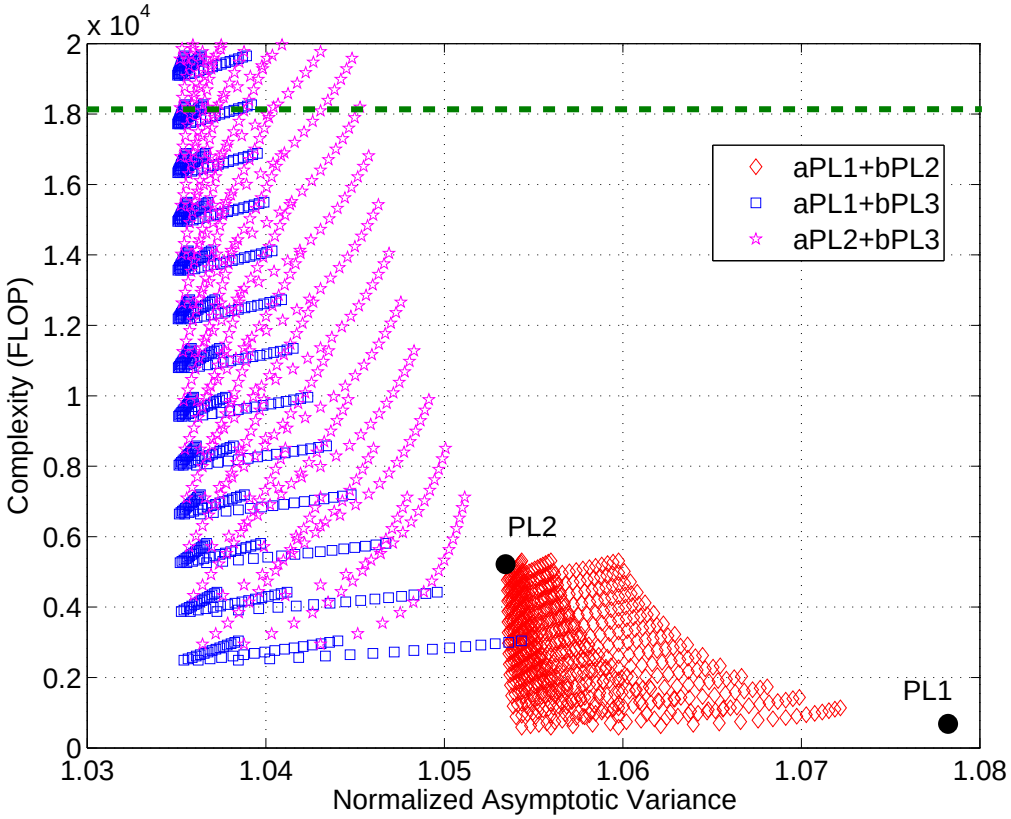

Figure 4: Computation-accuracy diagram for three SCL families:$\lambda_1\beta_1$PL1 + $\lambda_2(1-\beta_1)$PL2, $\lambda_1\beta_1$PL1 + $\lambda_2(1-\beta_1)$PL3, $\lambda_1\beta_1$PL2 + $\lambda_2(1-\beta_1)$PL3 (for multiple values of $\lambda_1, \lambda_2, \beta_1$) for the Boltzmann machine with 9 binary nodes. The pure policies PL1 and PL2 are indicated by black circles and the computational complexity of the full likelihood indicated by a dashed line (corresponding normalized asymptotic variance is 1). The PL3 pure policy is beyond the scale of the diagram. As the graph size increases, the computational cost increases dramatically, in particular for the full likelihood policy and to a lesser extent for the pseudo likelihood policy.

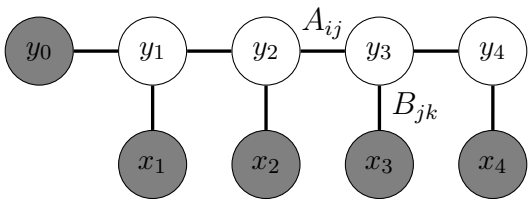

Figure 5: Graphical representation of a four token conditional random field (CRF). A, B are weight matrices and represent state-to-state transitions and state-to-observation outputs. Shading indicates the variable is conditioned upon while no shading indicates the variable is generated by the model.

As is common practice, we curtail overfitting through a L_2 regularizer, $\exp\{-(2n\sigma^2)^{-1}||\theta||_2^2\}$, which is strong when σ^2 is small and weak when σ^2 is large. We consider σ^2 a hyper-parameter and select it through cross-validation, unless noted otherwise.

Figure 6 shows the contour plots of train and test log-likelihood as a function of the SCL parameters: weight β and selection probability λ. The likelihood components were mixtures of full and pseudo ($|A_j| = 1$) likelihood (rows 1,3) and pseudo and 2nd order pseudo ($|A_j| = 2$) likelihood (rows 2,4). A_j identifies a set of labels corresponding to adjacent sentences over which the probabilistic query is evaluated. Results were averaged over 100 cross validation iterations with 50% train-test split. We used BFGS quasi-Newton method for maximizing the regularized SCL functions. The figure demonstrates how the train log-likelihood increases with increasing the weight and selection probability of full likelihood in rows 1,3 and of 2nd order pseudo likelihood in rows 2,4. This increase in train log-likelihood is also correlated with an increase in computational complexity as higher order likelihood components require more computation. Note however, that the test set behavior in the third and fourth rows shows an improvement in prediction accuracy associated with decreasing the influence of full likelihood in favor of pseudo likelihood. The fact that this happens for (relatively) weak regularization, $\sigma^2 = 10$, and indicates that lower order pseudo likelihood has a regularization effect which improves prediction accuracy when the model is not regularized enough. We have encountered this phenomenon in other experiments as well and we will discuss it further in the following subsections.

Figure 7 displays the complexity and negative log-likelihoods (left:train, right:test) of different SCL estimators, sweeping through λ and β, as points in a two dimensional space. The shaded area near the origin is unachievable as no SCL estimator can achieve high accuracy and low computation at the same time. The optimal location in this 2D plane is the curved boundary of the achievable region with the exact position on that boundary depending on the required solution of the computation-accuracy

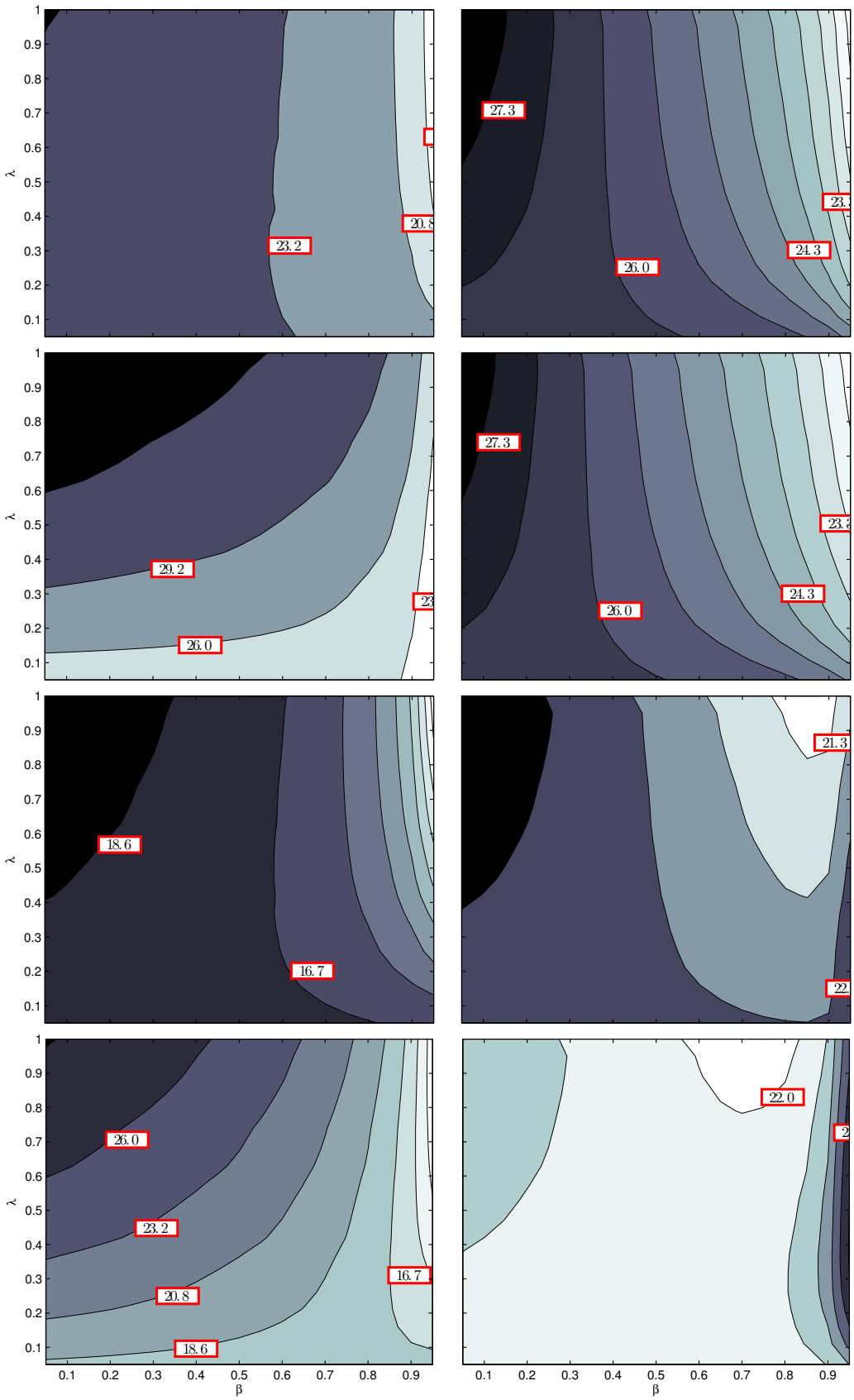

Figure 6: Train (left column) and test (right column) neg. log-likelihood contours for maximum SCL estimators for the CRF model. L_2 regularization, $\exp\{-(2n\sigma^2)^{-1}||\theta||_2^2\}$, parameters are $\sigma^2 = 1$ (rows 1,2) and $\sigma^2 = 10$ (rows 3,4). Rows 1,3 are stochastic mixtures of full (FL) and pseudo (PL1) log-likelihood components while rows 2,4 are PL1 and 2nd order pseudo likelihood (PL2). Most noteworthy, is the striking effect of the regularizer both in terms of other regularizers and the CoNLL-2000 dataset, Figure 16.

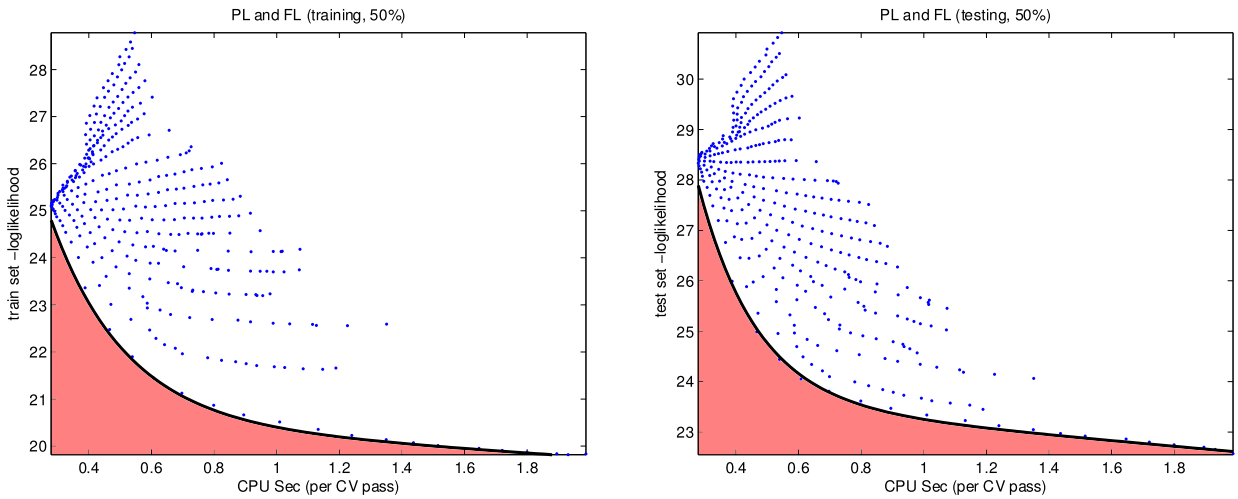

Figure 7: Scatter plot representing complexity and negative log-likelihood (left:train, right:test) of SCL functions for CRFs with L2 regularization parameter $\sigma^2 = 1/2$. The points represent different stochastic combinations of full and pseudo likelihood components. The shaded region represents impossible accuracy/complexity demands. Since the boundary of the obtainable region is empirical, the optimal beta always lies on this boundary. By varying λ, β we are able to smoothly span complexity (wall seconds) and accuracy.

tradeoff.

3.8.3 Text Chunking

This experiment consists of using sequential MRFs to divide sentences into "text chunks," i.e., syntactically correlated sub-sequences, such as noun and verb phrases. Chunking is a crucial step towards full parsing. For example,[6] the sentence:

> He reckons the current account deficit will narrow to only # 1.8 billion in September.

could be divided as:

> [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September].

where NP, VP, and PP indicate noun phrase, verb phrase, and prepositional phrase.

[6] Taken from the CoNLL-2000 shared task site, http://www.cnts.ua.ac.be/conll2000/chunking/.

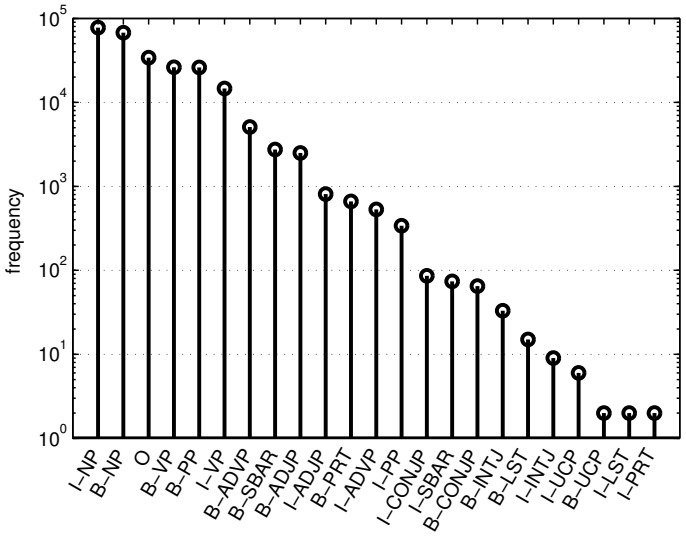

Figure 8: Label counts in CoNLL-2000 dataset. Phrases such as noun (NP), verb (VP), and prepositional (PP) are demarcated by a "begin" tag (B-) and an "inside" tag (I-). Non-phrase entities are denoted as "other" (O).

We used the publicly available CoNLL-2000 shared task dataset. It consists of labeled partitions of a subset of the Wall Street Journal (WSJ) corpus. Our training sets consisted of sampling 100 sentences without replacement from the the CoNLL-2000 training set (211,727 tokens from WSJ sections 15-18). The test set was the same as the CoNLL-2000 testing partition (47,377 tokens from WSJ section 20). Each of the possible 21,589 tokens, i.e., words, numbers, punctuation, etc., are tagged by one of 11 chunk types and an O label indicating the token is not part of any chunk. Chunk labels are prepended with flags indicating that the token begins (B-) or is inside (I-) the phrase. Figure 8 lists all labels and respective frequencies. In addition to labeled tokens, the dataset contains a part-of-speech (POS) column. These tags were automatically generated by the Brill tagger and must be incorporated into any model/feature set accordingly.

In the following, we explore this task using various SCL selection policies on two related, but fundamentally different sequential MRFs: Boltzmann chain MRFs and CRFs.

3.8.3.1 Boltzmann Chain MRF

Boltzmann chains are a generative MRF that are closely related to hidden Markov models (HMM). See [37] for a discussion on the relationship between Boltzmann chain MRFs and HMMs. We consider SCL components of the form $\mathsf{P}(X_2, Y_2|Y_1, Y_3)$, $\mathsf{P}(X_2, X_3, Y_2, Y_3|Y_1, Y_4)$ which we refer to as first and second order pseudo likelihood (with higher order components generalizing in a straightforward manner).

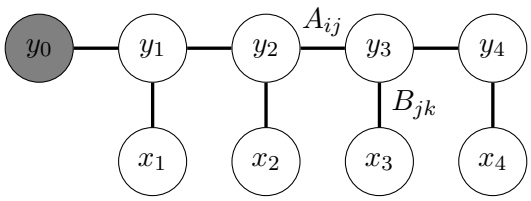

Figure 9: Graphical representation of a four token Boltzmann chain. A, B are weight matrices and represent preference in particular state-to-state transitions and state-to-feature emissions. Only the start state is conditioned upon while all others are generative.

The nature of the Boltzmann chain constrains our feature set to only encode the particular token present at each position, or time index. In doing so we avoid having to model additional dependencies across time steps and dramatically reduce computational complexity. Although SCL is precisely motivated by high treewidth graphs, we wish to include the full likelihood for demonstrative purposes–in practice, this is often not possible. Although POS tags are available we do not include them in these features since the dependence they share on neighboring tokens and other POS tags is unclear. For these reasons our time-sliced feature vector, x_i, has only a single-entry one and cardinality matching the size of the vocabulary (21,589 tokens).

As in Section 3.8.2, we control overfitting through a L_2 regularizer, $\exp\{-(2n\sigma^2)^{-1}||\theta||_2^2\}$, which is strong when σ^2 is small and weak when σ^2 is large. Here again we choose σ^2 via cross-validation unless otherwise noted. More often though, we show results for several representative σ^2 to demonstrate the roles of λ and β in $\hat{\theta}_n^{msl}$.

Figures 11 and 12 depict train and test negative log-likelihood, i.e., perplexity, for

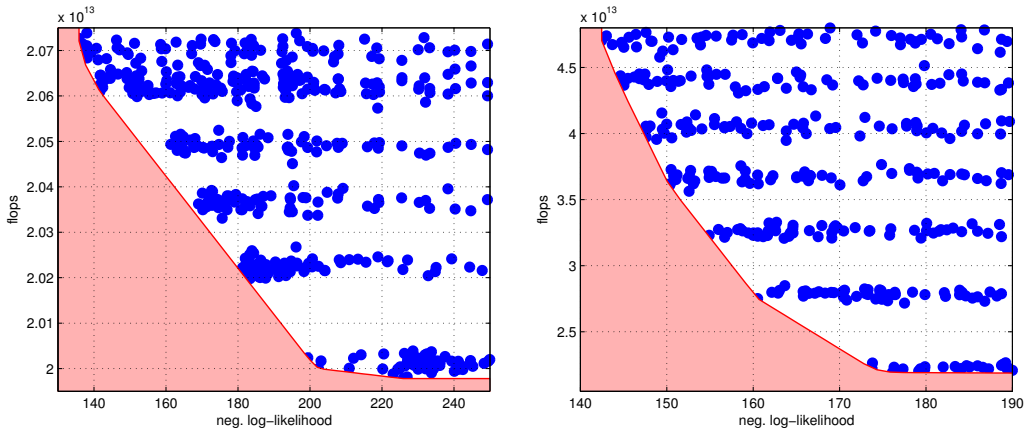

Figure 10: Accuracy and complexity tradeoff for the Boltzmann chain MRF with PL1/FL (left) and PL1/PL2 (right) selection policies. Each point represents the negative log-likelihood (perplexity) and the number of flops required to evaluate the composite likelihood and its gradient under a particular instantiation of the selection policy. The shaded region is the convex hull of the points and represents empirically unobtainable combinations of computational complexity and accuracy. Particularly interesting is the difference between policies and against the discriminative CRF, cf. Figure 16.

the SCL estimator $\hat{\theta}^{msl}_{100}$ with a pseudo/full likelihood selection policy (PL1/FL). As is our convention, weight β and selection probability λ correspond to the higher order component, in this case full likelihood. The lower order pseudo likelihood component is always selected and has weight $1-\beta$. As expected the test set perplexity dominates the train-set perplexity. As was the situation in Sec. 3.8.2, we note that the lower order component serves to regularize the full likelihood, as evident by the abnormally large σ^2.

We next demonstrate the effect of using a 1st order/2nd order pseudo likelihood selection policy (PL1/PL2). Recall, our notion of pseudo likelihood never entails conditioning on x, although in principle it could. Figures 13 and 14 show how the policy responds to varying both λ and β. Figure 10 depicts the empirical tradeoff between accuracy and complexity. Figure 15 highlights the effectiveness of the β heuristic. See captions for additional comments.

3.8.3.2 CRFs

Conditional random fields are the discriminative counterpart of Boltzmann chains (cf. Figures 5 and 9). Since x is not jointly modeled with y, we are free to include features with non-independence across time steps without significantly increasing the computational complexity. Here our notion of pseudo likelihood is more traditional, e.g., $\mathsf{P}(Y_2|Y_1, Y, 3, X_2)$ and $\mathsf{P}(Y_2, Y_3|Y_1, Y, 4, X_2, X_3)$ are valid 1st and 2nd order pseudo likelihood components.

We employ a subset of the features outlined in [45] which proved competitive for the CoNLL-2000 shared task. Our feature vector was based on seven feature categories, resulting in a total of 273,571 binary features (i.e., $\sum_i f_i(x_t) = 7$). The feature categories consisted of word unigrams, POS unigrams, word bigrams (forward and backward), and POS bigrams (forward and backward) as well as a stopword indicator (and its complement) as defined by [33]. The set of possible feature/label pairs is much larger than our set–we use only those features supported by the CoNLL-2000 dataset, i.e., those which occur at least once. Thus we modeled 297,041 feature/label pairs and 847 transitions for a total of 297,888 parameters. As before, we use the L_2 regularizer, $\exp\{-(2\sigma^2)^{-1}||\theta||_2^2\}$, which is strong when σ^2 is small and weak when σ^2 is large.

We demonstrate learning on two selection policies: pseudo/full likelihood (Figures 17 and 18) and 1st/2nd order pseudo likelihood (Figures 19 and 20). In both selection polices we note a significant difference from the Boltzmann chain, β has less impact on both train and test perplexity. Intuitively, this seems reasonable as the component likelihood range and variance are constrained by the conditional nature of CRFs. Figure 16 demonstrates the empirical accuracy/complexity tradeoff and Figure 21 depicts the effectiveness of the β heuristic. See captions for further comments.

3.8.4 Complexity/Regularization Win-Win

It is interesting to contrast the test log-likelihood behavior in the case of mild and stronger L_2 regularization. In the case of weaker or no regularization, the test log-likelihood shows different behavior than the train log-likelihood. Adding a lower order component such as pseudo likelihood acts as a regularizer that prevents overfitting. Thus, in cases that are prone to overfitting reducing higher order likelihood components improves both performance as well as complexity. This represents a win-win situation in contrast to the classical view where the MLE has the lowest variance and adding lower order components reduces complexity but increases the variance.

In Figure 6 we note this phenomenon when comparing $\sigma^2 = 1$ to $\sigma^2 = 10$ across the selection policies PL1/FL and PL1/PL2. That is, the weaker regularization and more restrictive selection policy, i.e., PL1/PL2, is able to achieve comparable test set perplexity.

For the text chunking experiments, we observe a striking win-win when using the Boltzmann chain MRF, Figures 11 and 13. Notice that as regularization is decreased (comparing from left to right), the contours are pulled closer to the x-axis. This means that we are achieving the same perplexity at reduced levels of computational complexity. The CRF however, only exhibits the win-win to a minor extent. We delve deeper into why this is might be the case in the following section.

3.8.5 λ, σ^2 Interplay

Throughout these experiments we fixed σ^2 and either swept over (λ, β) or used the heuristic to evaluate $(\lambda, \beta(\lambda))$. Motivated by the sometimes weak win-win (cf. Section 3.8.4) we now consider how the optimal σ^2 changes as a function of λ. In Figure 22 we used the β heuristic to evaluate train and test perplexity over a (λ, σ^2) grid. We used CRFs and the text chunking task as outlined in Section 3.8.3.2.

For the PL1/FL policy, we observe that for small enough λ the optimal σ^2, i.e.,

the σ^2 with smallest test perplexity, has considerable range. At some point there are enough samples of the higher-order component to stabilize the choice of regularizer, noting that it is still weaker than the optimal full likelihood regularizer. Conversely, the PL1/PL2 regularizer has an essentially constant optimal regularizer which is relatively much weaker.

As a result, we believe that the lack of win-win for the chunking CRF follows from two effects. In the case of the PL1/FL policy the contour plots are misleading since there is no single σ^2 that performs well across all $\lambda \in [0, 1]$. For the PL1/PL2 there is simply little change in regularization necessary across λ.

3.9 Discussion

The proposed estimator family facilitates computationally efficient estimation in complex graphical models. In particular, different (β, λ) parameterizations of the stochastic composite likelihood enables the resolution of the complexity-accuracy tradeoff in a domain and problem specific manner. The framework is generally suited for Markov random fields, including conditional graphical models and is theoretically motivated. When the model is prone to overfit, stochastically mixing lower order components with higher order ones acts as a regularizer and results in a win-win situation of improving test-set accuracy and reducing computational complexity at the same time.

It is interesting to note that the SCL framework may be generalized to random m-estimators beyond likelihood objects. That is, instead of a fixed m-function we may consider a linear combination of stochastic objects (appearing or not with some probability). Such estimators go beyond traditional m-estimator but may be analyzed using techniques similar to the ones developed in this chapter. Although not a random m-estimator, the work of [22] borrows SCL concepts to facilitate budgeted semi-supervised learning. This too would benefit from a random m-estimator interpretation and indeed many machine learning tasks may fit nicely into such a framework.

The SCL framework may be useful for a wide variety of intractable graphical models. Besides the examples presented here, it may be particularly suited for large scale models from statistical physics, exponential random graph models, and models from computational biology. A particularly nice feature is that the above computation may be trivially parallelized thus leading to effective computation on large clusters and cloud computing.

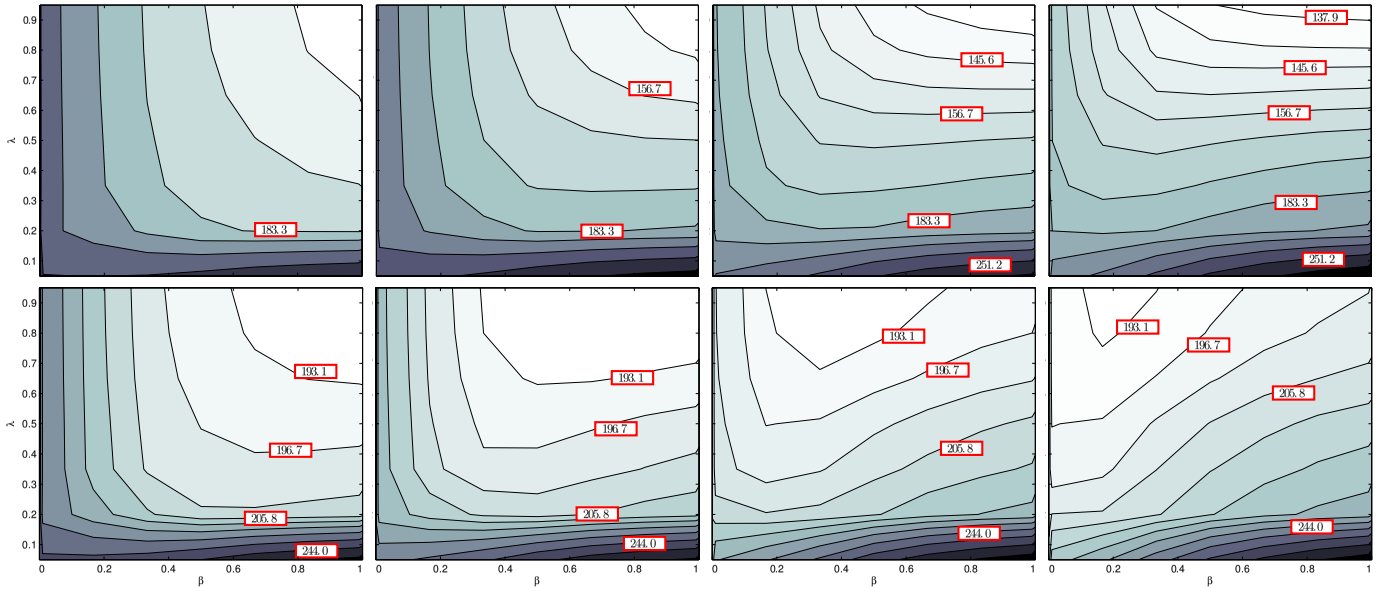

Figure 11: Train set (top) and test set (bottom) negative log-likelihood (perplexity) for the Boltzmann chain MRF with pseudo/full likelihood selection policy (PL1/FL). The x-axis, β, corresponds to relative weight placed on FL and and the y-axis, λ, corresponds to the probability of selecting FL. PL1 is selected with probability 1 and weight $1 - \beta$. Contours and labels are fixed across columns. Results averaged over several cross-validation folds, i.e., resampling both the train set and the PL1/FL policy. Columns from left to right correspond to weaker regularization, $\sigma^2 = \{500, 1000, 2500, 5000\}$. The best achievable test set perplexity is about 190.

Unsurprisingly the test set perplexity dominates the train set perplexity at each σ^2 (column). For a desired level of accuracy (contour) there exists a computationally favorable regularizer. Hence $\hat{\theta}_n^{msl}$ acts as both a regularizer and mechanism for controlling accuracy and complexity.

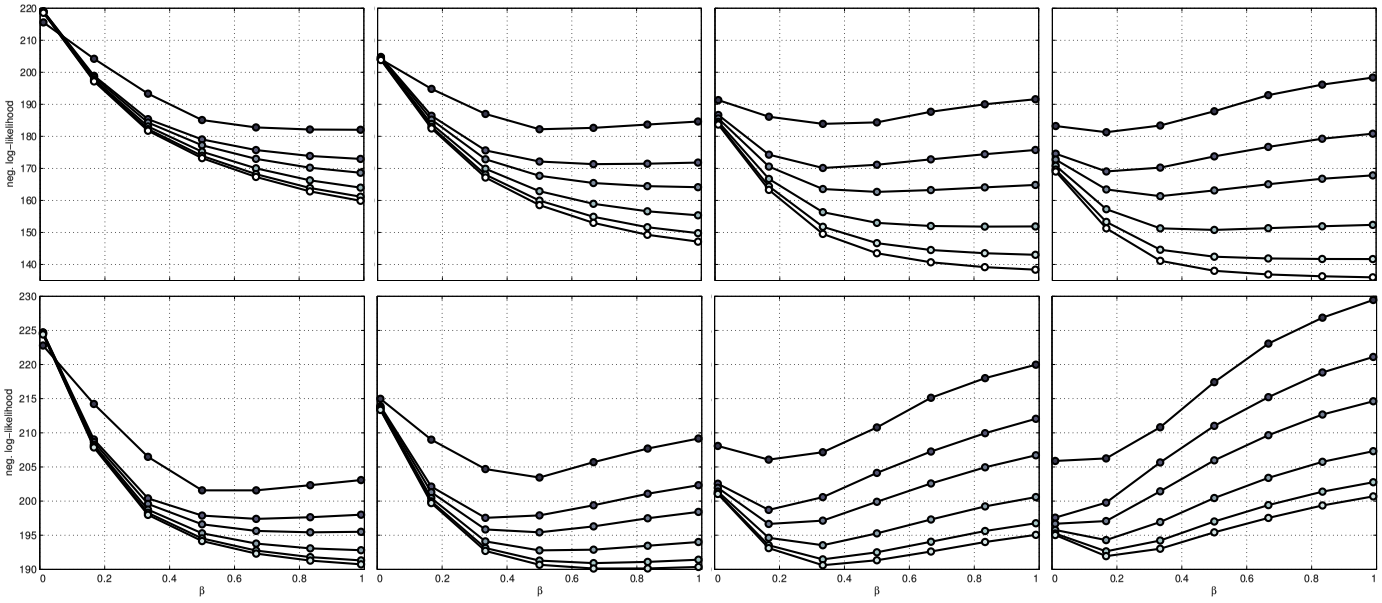

Figure 12: Train set and test set perplexities for the Boltzmann chain MRF with PL1/FL selection policy (see above layout description). The x-axis is again β and the y-axis perplexity. Lighter shading indicates FL is selected with increasing frequency. Note that as the regularizer is weakened the range in perplexity spanned by λ increases and the lower bound decreases. This indicates that the approximating power of $\hat{\theta}_n^{msl}$ increases when unencumbered by the regularizer and highlights its secondary role as a regularizer.

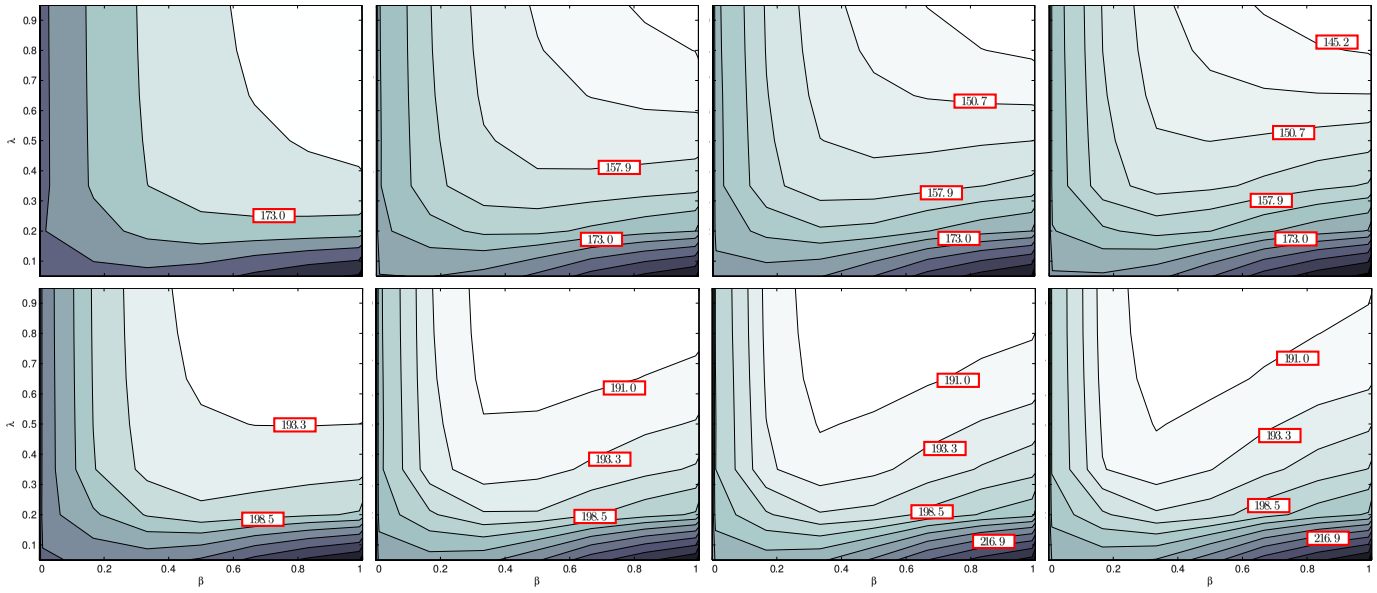

Figure 13: Train set (top) and test set (bottom) perplexity for the Boltzmann chain MRF with 1st/2nd order pseudo likelihood selection policy (PL1/PL2). The x-axis corresponds to PL2 weight and the y-axis the probability of its selection. PL1 is selected with probability 1 and weight $1 - \beta$. Columns from left to right correspond to $\sigma^2 = \{5000, 10000, 12500, 15000\}$. See Figure 11 for more details. The best achievable test set perplexity is about 189.5.

In comparing these results to PL1/FL, we note that the test set contours exhibit less perplexity for larger areas. In particular, perplexity is lower at smaller λ values, meaning a computational saving over PL1/FL at a given level of accuracy.

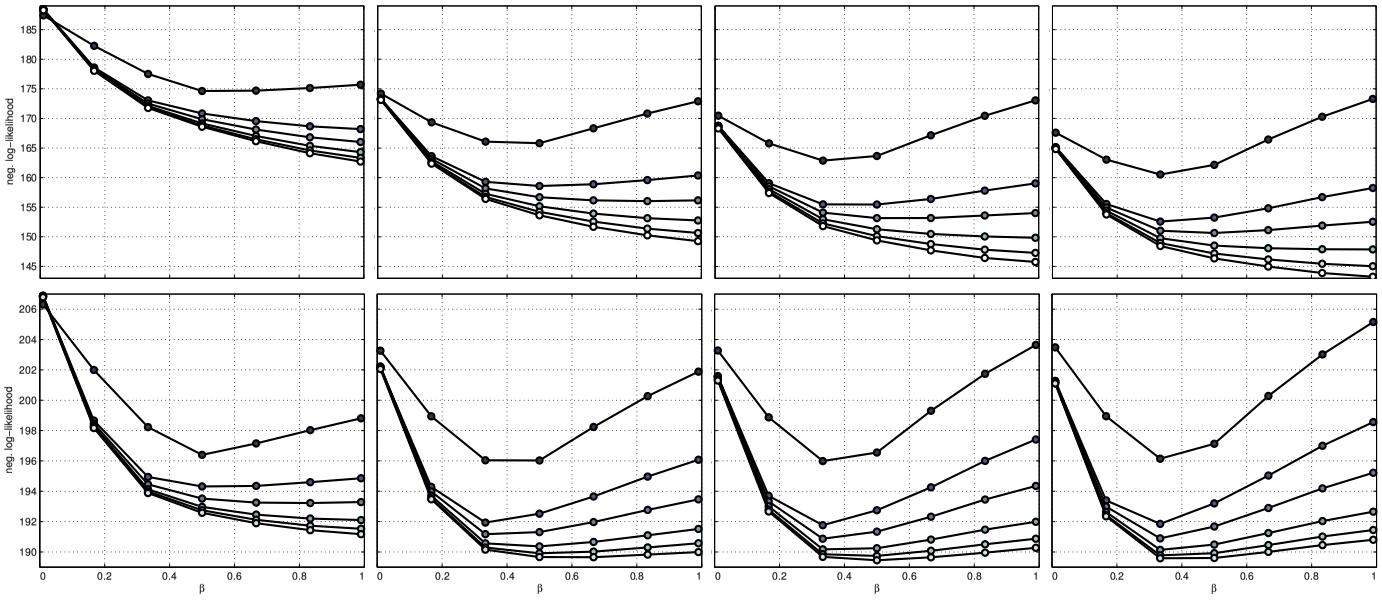

Figure 14: Train (top) and test (bottom) perplexities for the Boltzmann chain MRF with PL1/PL2 selection policy (x-axis:PL2 weight, y-axis:perplexity; see above and previous).

PL1/PL2 outperforms PL1/FL test perplexity at $\sigma^2 = 5000$ and continues to show improvement with weaker regularizers. This is perhaps surprising since the previous policy includes FL as a special case, i.e., $(\lambda, \beta) = (1, 1)$. We speculate that the regularizer's indirect connection to the training samples precludes it from preventing certain types of overfitting. See Sec. 3.8.4 for more discussion.

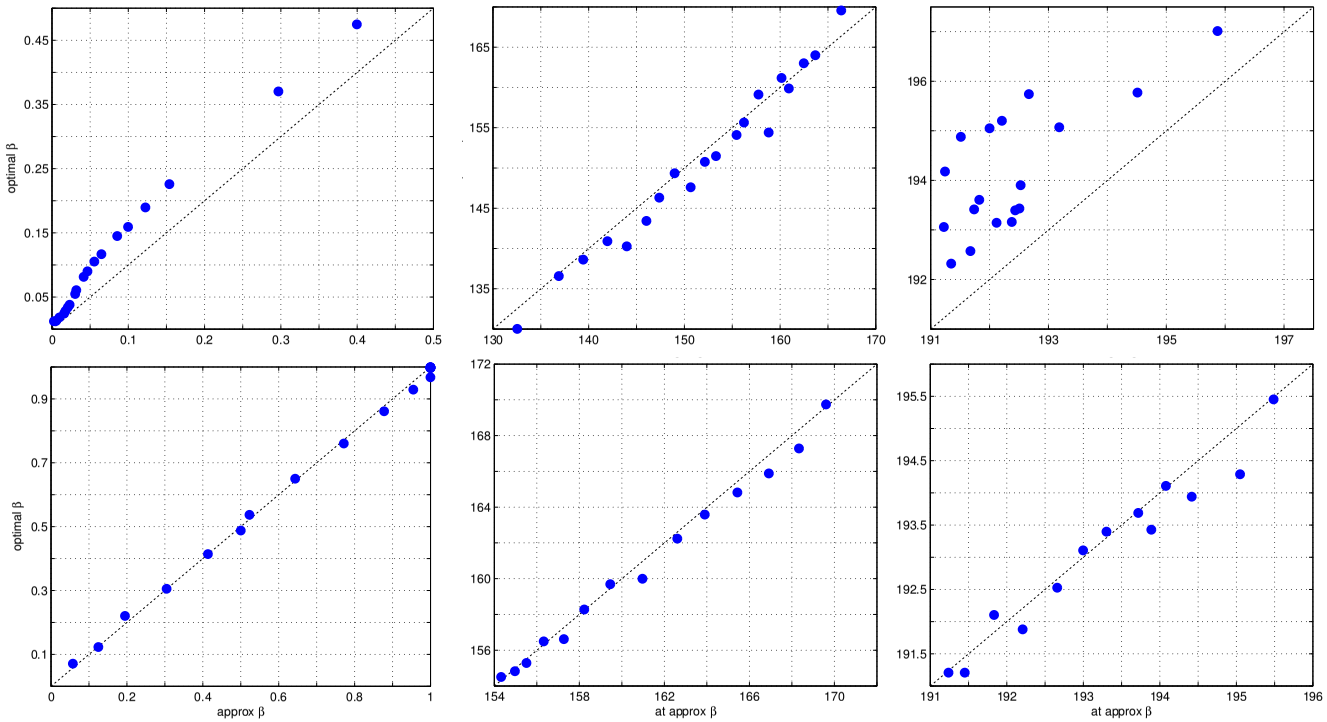

Figure 15: Demonstration of the effectiveness of the β heuristic, i.e., using $\hat{\theta}^{msl}$ as a plug-in estimate for θ_0 to periodically re-estimate β during gradient descent. Results are for the Boltzmann chain with PL1/FL (top) and PL1/PL2 (bottom) selection policies. The x-axis is the value at the heuristically found β and the y-axis the value at the optimal β. The optimal β was found by evaluating over a β grid and choosing that with the smallest train set perplexity. The first column depicts the best performing β against the heuristic β. The second and third columns depict the training and testing perplexities (resp.) at the best performing β and heuristically found β. For all three columns, we assess the effectiveness of the heuristic by its nearness to the diagonal (dashed line).

For the PL1/PL2 policy the heuristic closely matched the optimal (all bottom row points are on diagonal). The heuristic out-performed the optimal on the test set and had slightly higher perplexity on the training set. It is a positive result, albeit somewhat surprising, and is attributable to either coarseness in the grid or improved generalization by accounting for variability in $\hat{\theta}^{msl}$.

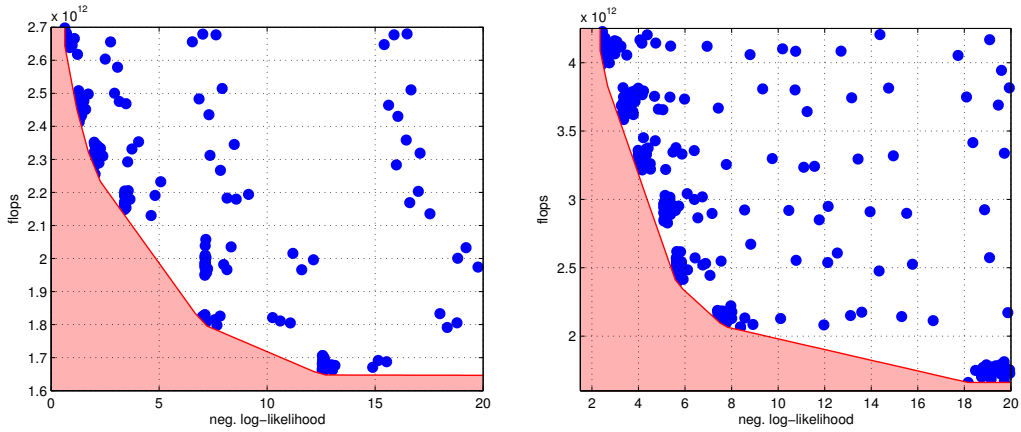

Figure 16: Accuracy and complexity tradeoff for the CRF with PL1/FL (left) and PL1/PL2 (right) selection policies. Each point represents the negative log-likelihood (perplexity) and the number of flops required to evaluate the composite likelihood and its gradient under a particular instance of the selection policy. The shaded region is the convex hull of the points and represents empirically unobtainable combinations of computational complexity and accuracy. σ^2. Particularly interesting is the difference between policies and against the generative Boltzmann chain, cf. Figure 10.

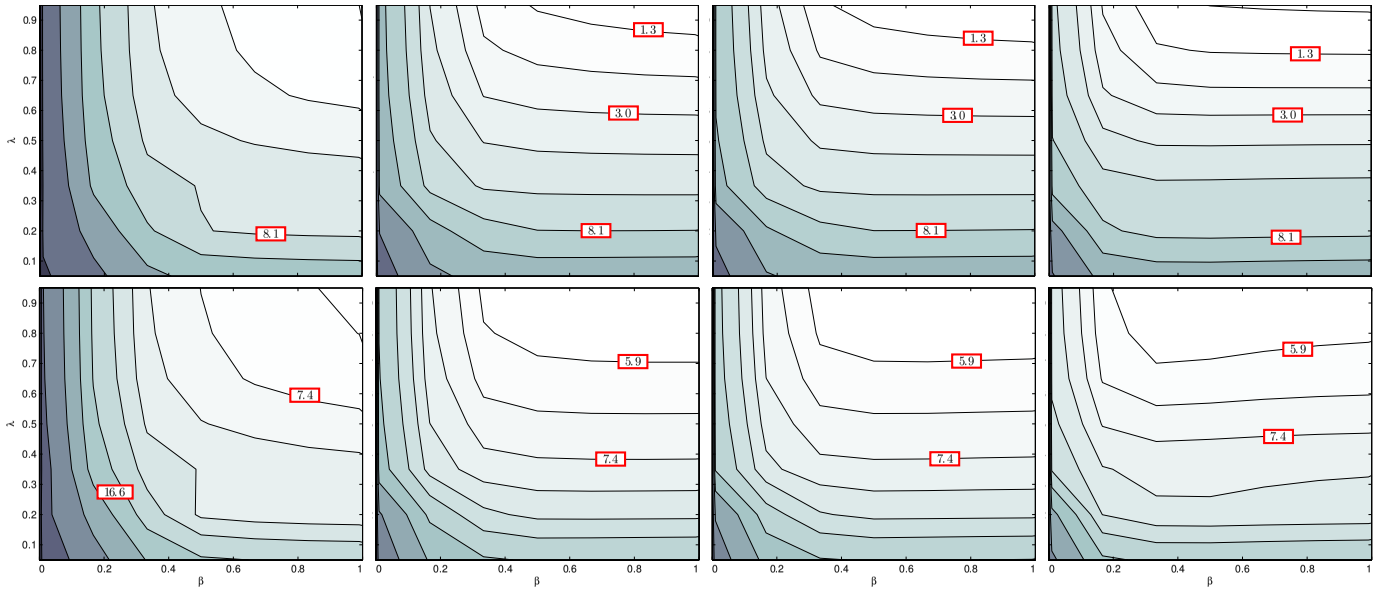

Figure 17: Train set (top) and test set (bottom) perplexity for the CRF with pseudo/full likelihood selection policy (PL1/FL). The x-axis corresponds to FL weight and the y-axis the probability of its selection. PL1 is selected with probability 1 and weight $1 - \beta$. Columns from left to right correspond to $\sigma^2 = \{5000, 10000, 12500, 15000\}$. See Figure 11 for more details. The best achievable test set perplexity is about 5.5.

Although we cannot directly compare CRFs to its generative counterpart, we observe some strikingly different trends. It is immediately clear that the CRF is less sensitive to the relative weighting of components than is the Boltzmann chain. This is partially attributable to a smaller range of the objective–the CRF is already conditional hence the per-component perplexity range is reduced.

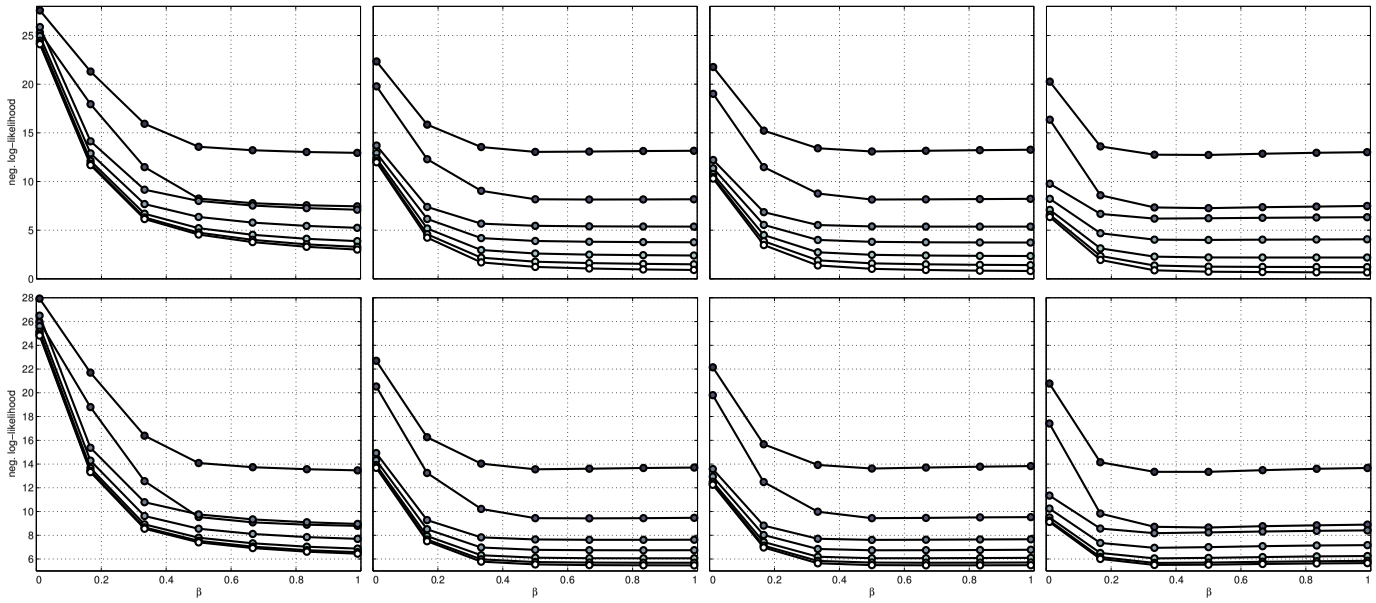

Figure 18: Train (top) and test (bottom) perplexities for a CRF with PL1/FL selection policy (x-axis:FL weight, y-axis:perplexity; see above and Fig. 12).

Perhaps more evidently here than above, we note that the significance of a particular β is less than that of the Boltzmann chain. However, for large enough σ^2, the optimal $\beta \neq 1$. This indicates the dual role of PL1 as a regularizer. Moreover, the left panel calls attention to the interplay between β, λ, and σ^2. See Sec. 3.8.5 for more discussion.

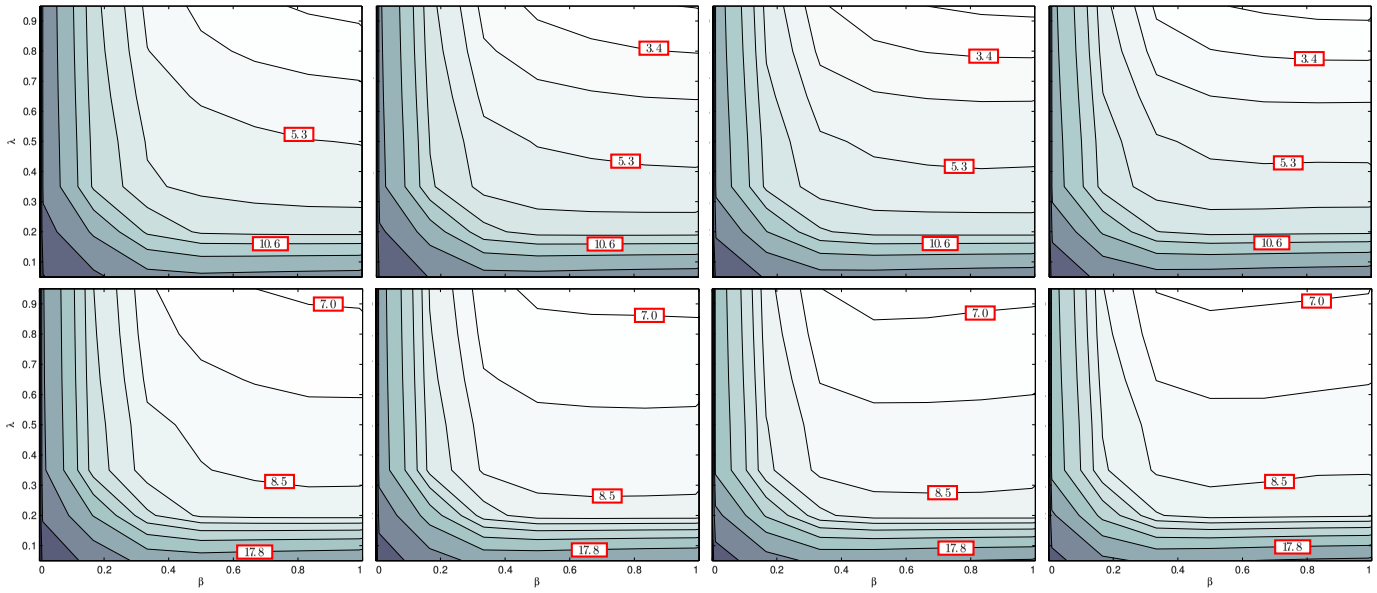

Figure 19: Train set (top) and test set (bottom) perplexity for a CRF with 1st/2nd order pseudo likelihood selection policy (PL1/PL2). The x-axis, β, represents the relative weight placed on PL2 and the y-axis, λ, the probability of selecting PL2. PL1 is selected with probability 1. Columns from left to right correspond to weaker regularization, $\sigma^2 = \{10000, 20000, 30000, 40000\}$. See Figure 17 for more details.

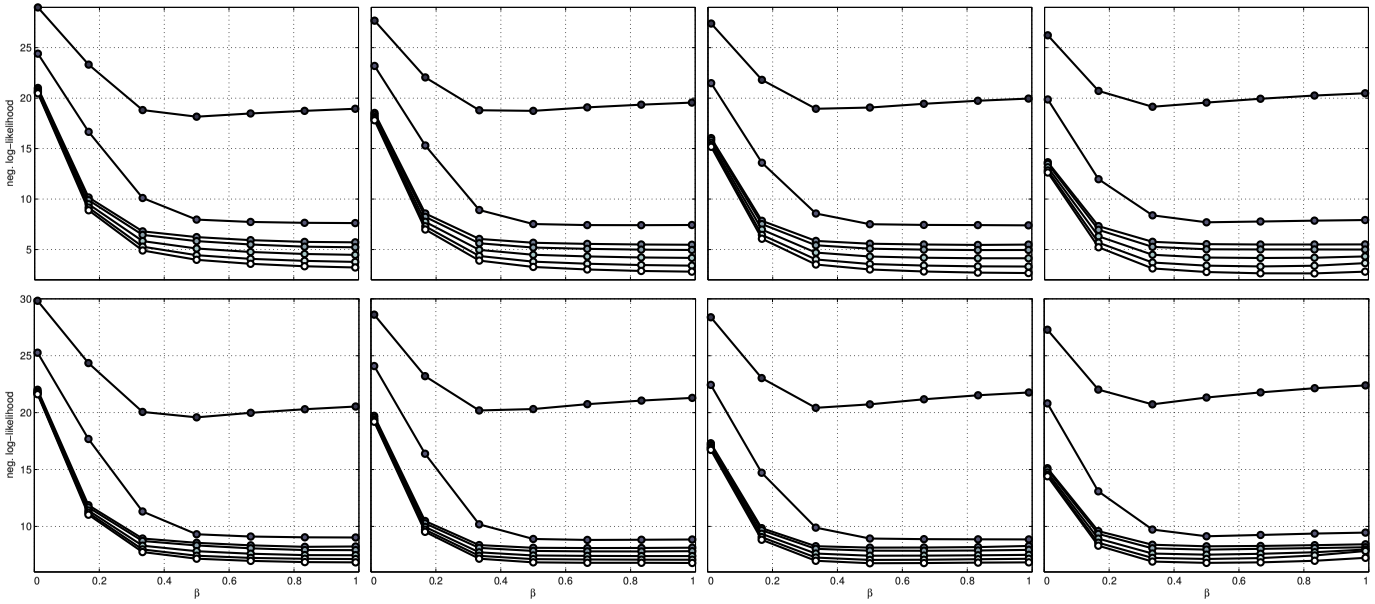

Figure 20: Train (top) and test (bottom) perplexities for a CRF with PL1/PL2 selection policy (x-axis:PL2 weight, y-axis:perplexity; see above and Fig. 12).

Although increasing λ only brings minor improvement to both the training and testing perplexities, it is worth noting that the test perplexity meets that of the PL1/FL. Still though, the overall lack of resolution here suggests that smaller values of λ would better span a range of perplexities and at reduced computational cost.

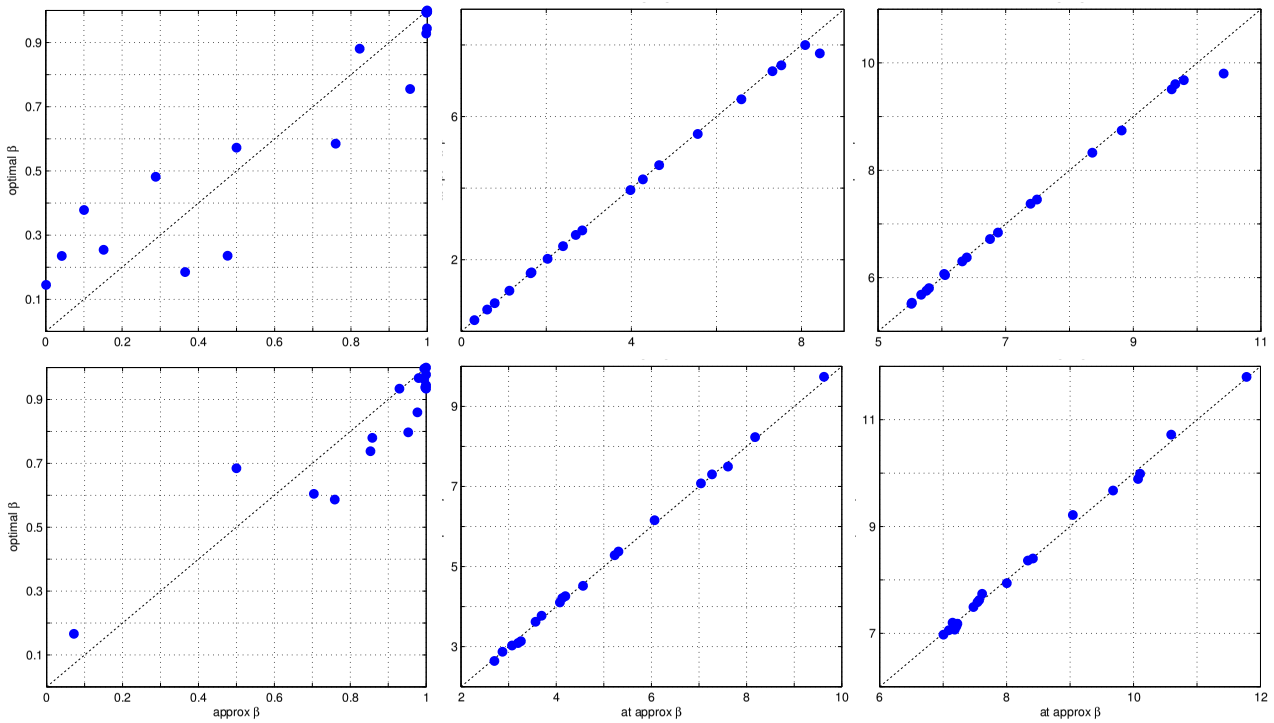

Figure 21: Demonstration of the effectiveness of the β heuristic. Results are for the CRF with PL1/FL (top) and PL1/PL2 (bottom) selection policies. The x-axis is the value at the heuristically found β and the y-axis the value at the optimal β. The first column depicts the best performing β against the heuristic β. The second and third columns depict the training and testing perplexities (resp.) at the best performing β and heuristically found β. For all three columns, we assess the effectiveness of the heuristic by its nearness to the diagonal (dashed line). See Fig. 15 for more details.

The optimal and heuristic β match train and test perplexities for both policies. The actual β value however does not seem to match as well as the Boltzmann chain. However, if we note the flatness of the β grid (cf. Fig. 18 and 20) this result is unsurprising and can be disregarded as an indication of the heuristic's performance.

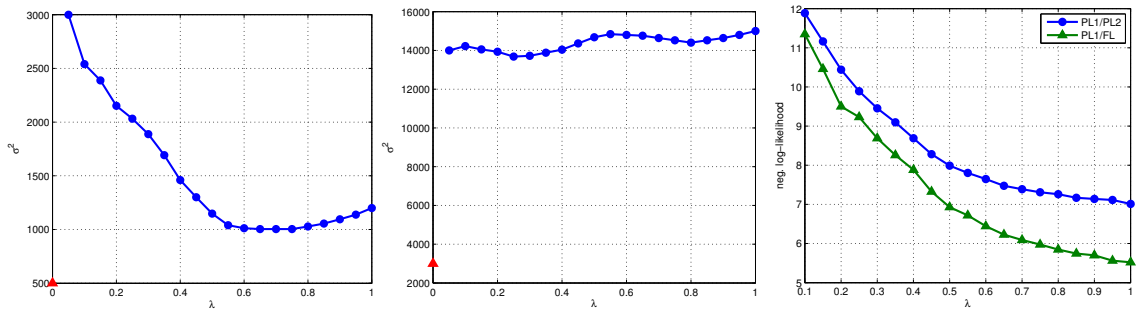

Figure 22: Optimal regularization parameter as a function of $(\lambda, \hat{\beta}(\lambda))$ for PL1/FL (left) and PL1/PL2 (center) CRF selection policies. In the left figure, PL1/FL, λ represents the probability of including FL into the objective. A few FL samples add uncertainty to the objective thus a weaker regularizer is preferable. As more FL samples are incorporated, this effect diminishes but still acts to regularize since the full likelihood (only) best regularization is $\sigma^2 = 500$ (red triangle). The center figure, PL1/PL2, exhibits only a minor change as λ (the probability of incorporating PL2) is increased. It is however, best served by a much weaker regularizer than PL2 alone (red triangle).

The right figure depicts the test-set perplexity as a function of λ using the optimal σ^2 (small λ values were clipped as their performance is quite poor). Note that the perplexity is lowest when both components are always selected ($\lambda = 1$) and that the PL1/FL policy outperforms the PL1/PL2 policy as expected.

CHAPTER IV

COMPUTATIONAL COST: LEARNING HIDDEN MRFS

As in the previous chapter we shall continue to concern ourselves with the challenges present learning parameters of high tree-width undirected graphical models. As before the techniques developed presently are of practical significance only when the efficient inference algorithms do not exist. This chapter differs from the previous by specializing the analysis to those models which are naturally regarded as being a marginal of some joint distribution. This is essential for maximum likelihood type inference when the model has unobserved or hidden random variables.

4.1 Introduction

We begin by recalling the notion developed in the previous chapter and add to it as needed.

Let $\{p_\theta : \theta \in \Theta\}$ be a parametric family with members defined on an m-dimensional measure space $(\mathcal{X}, \mathcal{A}, \mu)$, and let (Θ, d) be an r-dimensional metric space $(m, r < \infty)$. Here again we will concern ourselves with identifying the index $\theta \in \Theta$ such that p_θ most closely resembles some distribution of nature p.[1] The resemblance shall be characterized through the intermediary p_n, i.e., the empirical distribution constructed from an iid sequence $\{x^{(i)}\}_1^n$ where $x^{(1)} \sim p$. We refer to the sequence as the training data or dataset.it by D_n.

[1]For the entirety of this chapter we assume that the measure μ is σ-finite and that the Radon-Nikodym derivatives of all distribution functions, e.g., P_θ, P, Q, exist under a common measure, e.g., $\mathrm{d}P_\theta(x) = p_\theta(x)\mathrm{d}\mu(x)$. When unambiguous, we use the terms "distribution function" and "probability function" interchangeably, and often simply refer to the function in question as a distribution.

4.2 Additional Notation

An *exponential family* $\mathcal{E} = (T, \mu, \Theta)$ is a parametric family whose members are linear-energy Markov random fields (47), i.e., $E_\theta(X) = -\theta^\mathsf{T} T(X)$. The function $T : \mathcal{X} \to \mathbb{R}^r$ is non-random and μ is a base measure.

The *Kullback–Leibler divergence*, from p to q is defined as,

$$\mathrm{D}_{\mathsf{KL}}\big(p \,\big\|\, q\big) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \,\mathrm{d}\mu(x),$$

where P being absolutely continuous with respect to Q justifies the convention $0 \log 0 = 0$. When considering members of $\{p_\theta : \theta \in \Theta\}$ we often write $\mathrm{D}_{\mathsf{KL}}\big(p_\eta \,\big\|\, p_\theta\big)$ as $\mathrm{D}_{\mathsf{KL}}\big(\eta \,\big\|\, \theta\big)$.

The conditional KL-divergence should be interpreted as the expected KL-divergence, i.e., for a particular $a \in \mathcal{A}$ such that $p(a), p(a^c) > 0$ and $q(a), q(a^c) > 0$,

$$\begin{aligned}\mathrm{D}_{\mathsf{KL}}\big(p(X|A) \,\big\|\, q(X|A)\big) &= p(a)\, \mathrm{D}_{\mathsf{KL}}\big(p(X|a) \,\big\|\, q(X|a)\big) \\ &\quad + p(a^c)\, \mathrm{D}_{\mathsf{KL}}\big(p(X|a^c) \,\big\|\, q(X|a^c)\big)\,,\end{aligned}$$

where $X|a$ indicates the right-hand side should be interpreted as $a \mapsto \mathrm{D}_{\mathsf{KL}}\big(p(X|a) \,\big\|\, q(X|a)\big)$. Strict adherence to the convention that the event A is a random variable and a is a random variate makes the distinction between left- and right-hand-side clear.

The significance of KL-divergence in the following discussion follows from the same logic used in the previous chapter. Fundamentally, we wish to exploit its relationship to the likelihood function as well as the Gibbs Inequality, which we now re-state.

Lemma 3. *Kullback–Leibler divergence of p and q defined on measure space $(\mathcal{X}, \mathcal{A}, \mu)$ is non-negative for all p, q and zero if and only if p is identically q.*

Proof.

$$-\mathrm{D}_{\mathsf{KL}}\big(p \,\big\|\, q\big) = \mathsf{E}_p \log \frac{q(X)}{p(X)} \leq \log \mathsf{E}_p \frac{q(X)}{p(X)} = \log 1 = 0 \tag{45}$$

where the bound and its unique tightness follow from Jensen's inequality. □

Since maximizing the likelihood criterion $\theta \mapsto \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x^{(i)})$ is equivalent to maximizing $-\,\mathrm{D}_{\mathrm{KL}}\big(p_n \,\big\|\, p_\theta\big)$, the Gibbs inequality is an essential property for establishing conditions for the theoretical convergence of the estimator.

4.3 Hidden MRF

Unlike the previous chapter however, we now imagine that only a portion of each sample is observed. We say "imagine" because it is often the case that the "unobserved data" has no statistical meaning. A more general treatment would regard the "unobserved data" as merely auxiliary (non-random) variables. However, we adhere to this analogy as it provides a rich conceptual device for understanding a particular class of MRFs which we describe presently.

In accordance to the "missing data" analogy, we shall interpret the sample space $\mathcal{X}$ as a Cartesian product of spaces corresponding to observed and unobserved data, i.e., $\mathcal{X} = \mathcal{Y} \times \mathcal{Z}$ and $\mu(x) = \mu(y)\mu(z)$. Such an assumption implies that all samples have the same configuration of missing and non-missing elements. That the measure μ should factorize is a matter of convenience (although it is typical) and is unimportant to the presentation. It is appropriate to refer to $\{x^{(i)}\}_1^n$ as the complete dataset and $\{y^{(i)}\}_1^n$ as the observed dataset.

That this framework still corresponds to a Markov random field is apparent from establishing some additional notation. To simplify subsequent presentation, we begin by writing an MRF in the log-domain, i.e.,

$$p_\theta(x) = \prod_{c \in C} \phi_c(x_c; \theta) \tag{46}$$

$$\stackrel{\text{def}}{=} \exp\left\{-E_\theta(x) + A_\theta\right\} \mu(x), \tag{47}$$

where, as before, the set of cliques C is an appropriate set-of-sets and the clique potential ϕ_c is a non-negative function of sub-vector x_c. In the log-domain, the function $A_\theta = -\log Z_\theta$ serves as normalization and the function $E_\theta(x) = -\sum_{c \in C} \log \phi_c(x_c)$,

or "energy function," conveys the structural relationship between data elements (by way of clique potentials). The product measure $\mu(x)$ typically serves to control the support of the family; it cannot depend on θ and is an implied consequence of the construction of each clique potential.

Working with "energy" rather than directly manipulating probabilities simplifies the analysis by eliminating at least one source of non-linearity. The terminology however should be seen as equivalent; the energy formulation is merely a (negative) log-domain representation of an MRF. In addition to simplifying analysis, this terminology provides a reasonable descriptive vocabulary with a rich history in the physics community.

Index the observed portion of x as $y(x)$, or where unambiguous, simply y and similarly denote the unobserved portion of x by $z(x)$ or simply z. Under this notation, a hidden Markov random field may be defined as,

$$p_\theta(y) = \exp\{-F_\theta(y) + A_\theta\}\mu(y), \text{ where,} \tag{48}$$

$$F_\theta(y) = -\log \int_{\{y\}\times\mathcal{Z}} \exp\{-E_\theta(x)\}\, \mathrm{d}\mu(x), \tag{49}$$

$$A_\theta = -\log \int_{\mathcal{Y}} \exp\{-F_\theta(y)\}\, \mathrm{d}\mu(y). \tag{50}$$

The function F_θ is known as the variational free-energy and, for any fixed y, bears close resemblance to the normalization term of a standard MRF. Written in this manner, we can understand the hidden MRF as simply an MRF with a particular form of energy function, viz. the total energy in a particular cell of a fixed partitioning of the sample space.

4.3.0.1 Learning Challenge

Unfortunately, this construction often precludes the computational advantages of a likelihood based on conditional components. For example, consider the following

observed-data full-conditional,

$$p_\theta(y_j|y_{-j}) = \exp\left\{-F_\theta(y) - A_\theta(y_{-j})\right\}\mu(y), \text{ where,}$$
$$A_\theta(y_{-j}) = -\log\int_{\mathcal{Y}_j\times\{y_{-j}\}}\exp\left\{-F_\theta(y))\right\}\,\mathrm{d}\mu(y).$$

The computation of $A_\theta(y_{-j})$ is reduced to integration of only one dimension of $\mathcal{Y}$, however it remains necessary to integrate over all dimensions of $\mathcal{Z}$ (for *every point* in $\mathcal{Y}_j \times \{y_{-j}\}$).

To address the inability of the SCL framework to provide a computationally simple approximation of $F_\theta(y)$, we examine a alternative approach known as Monte Carlo Expectation. We show that MCEM and SCL possess orthogonal strengths and weaknesses, and describe how a MCEM, SCL hybrid is a particularly compelling framework for coping with the computational burden of $F_\theta(y)$.

4.4 EM, Importance Sampling, & SCL

In this section we recall Expectation Maximization (EM) and its Monte Carlo Markov Chain (MCMC) variant, Monte Carlo EM (MCEM). We then develop an MCEM+SCL hybrid algorithm and show how it spans the accuracy/cost tradeoff for latent variable MRFs.

4.4.1 Expectation Maximization

Expectation maximization is an immensely popular technique for coping with objectives of the form,

$$\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log\int_{\{y^{(i)}\}\times\mathcal{Z}} p_\theta(y,z)\,\mathrm{d}\mu(x),$$

by deriving a variational objective and solving this in a coordinate-wise manner. The surrogate can be derived as follows,

$$\begin{aligned}
\ell_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \log \int_{\{y^{(i)}\}\times\mathcal{Z}} p_\theta(y,z)\frac{q(z|y)}{q(z|y)}\,\mathrm{d}\mu(x) \\
&\geq \int_{\mathcal{X}} p_n(y)q(z|y)\log\frac{p_\theta(y,z)}{q(z|y)}\,\mathrm{d}\mu(x) \qquad (51)\\
&= \ell_n(\theta) - \int_{\mathcal{Y}} \mathrm{D}_{\mathsf{KL}}\big(q(Z|y)\,\big\|\,p_\theta(Z|y)\big)\,\mathrm{d}P_n(y) \qquad (52)\\
&\triangleq Q_n(\theta, q)
\end{aligned}$$

for some distribution q in a set of distributions $\mathcal{Q}$ which has positive support on $\mathcal{Z}$. Equation (51) follows from application of Jensen's inequality and (52) makes clear when this inequality is tight (since Gibbs inequality asserts $\mathrm{D}_{\mathsf{KL}}\big(p\,\big\|\,q\big) = 0$ iff $p \equiv q$). Henceforth, we assume $\mathcal{Q} = \{p_\theta(Y|X) : \theta \in \Theta\}$ and may therefore write $Q_n(\theta, q_\xi)$ as $Q_n(\theta,\xi)$ without ambiguity (cf. Thm. 6).

By Gibbs inequality and (52), $\ell_n(\theta) = Q_n(\theta,\theta) = \sup_{\xi\in\Theta} Q_n(\theta,\xi)$, and we may equivalently view the MLE as,

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta}\max_{\xi\in\Theta} Q_n(\theta,\xi).$$

That such a formulation should be computationally favorable follows from alternating the respective maximizations, i.e., one may solve $Q_n(\theta,\xi)$ coordinate-wise in θ and ξ. A general EM algorithm can thus be stated as a the following three-step procedure (given some initial $\theta^{(0)}$),

$$\begin{aligned}
&\text{E-step:} && \xi^{(t)} \leftarrow \arg\max_{\xi\in\Theta} Q_n(\theta^{(t-1)},\xi)\\
&\text{M-step:} && \theta^{(t)} \leftarrow \arg\max_{\theta\in\Theta} Q_n(\theta,\xi^{(t)})\\
&\text{Loop:} && t \leftarrow t+1. \qquad (53)
\end{aligned}$$

When $\mathcal{Q} = \{p_\theta : \theta\in\Theta\}$, the E-step has the analytical solution $\xi^{(t)} = \theta^{(t-1)}$, again by the properties of KL-divergence. That this algorithm converges can be understood

as a consequence of the ascent property, i.e.,

$$\ell_n(\theta^{(t-1)}) = Q_n(\theta^{(t-1)}, \theta^{(t-1)}) \leq Q_n(\theta^{(t)}, \theta^{(t-1)}) \leq \ell_n(\theta^{(t)}), \tag{54}$$

where each (in)equality follows from the respective majorizing properties of Q_n,

1. $\ell_n(\theta^{(t-1)}) = Q_n(\theta^{(t-1)}, \theta^{(t-1)})$,
2. $\theta^{(t)} \in \{\theta \in \Theta : Q_n(\theta, \theta^{(t-1)}) \geq Q_n(\theta^{(t-1)}, \theta^{(t-1)})\}$,
3. $\ell_n(\theta) \geq Q_n(\theta, \theta^{(t-1)})$.

Hence EM can be understood as an iterative approach for maximizing a lower bound of ℓ_n which is tight at the previous iteration's maximizer. [30] The statistical convergence of the sequence $\{\theta^{(t)}\}$ to the population maximizer $\theta_0 = \arg\min_\theta \mathrm{D}_{\mathrm{KL}}\big(p \,\big\|\, p_\theta\big)$ is a more delicate issue and is the subject of [57]. By examining the sequence as an intersection of point-to-set maps, Wu shows that the sequence is indeed statistically consistent with the points of local extrema under the population distribution.

Informally we can understand the statistical convergence as being a consequence of the (assumed) smoothness of Q_n and the property that $\ell_n(\theta) = Q_n(\theta, \theta)$, for which Proposition 1 ensures consistency.

That this approach is named EM rather than "Alternating Maximization" is a consequence of the assumptions made during its original formulation.[20] That is, assuming $\{p_\theta : \theta \in \Theta\}$ is an exponential family, the surrogate optimization objective may be rewritten as,

$$Q_n(\theta, \theta^{(t-1)}) = \mathsf{E}_n \, \mathsf{E}\left[\log p_\theta(Y, Z) \middle| \theta^{(t-1)}, Y\right], \tag{55}$$

where E_n denotes the expectation under the empirical distribution. Hence each iteration entails maximization of an expectation parametrized by the previous iteration's maximizer.

4.4.2 Monte Carlo EM

Like the stochastic composite likelihood framework, expectation maximization fundamentally rests on the assumption that computing the expectation over the posterior $q(Z|Y)$ is computationally simple (cf. (51)). In situations where this is not the case, we can expect MCMC techniques to offer reasonable alternative methods since the problematic term is already an expectation. This is precisely the intuition which underlies Monte Carlo EM (MCEM) algorithms [56].

We begin by observing that, under the assumption that the surrogate family $\mathcal{Q} = \{p_\theta : \theta \in \Theta\}$, the E-step of (53) is maximized by $\xi^{(t)} = \theta^{(t-1)}$ by the properties of KL-divergence. Regarding the M-step, we note that the $-\log p_{\theta^{(t-1)}}$ is constant in θ and its omission does not affect the maximizer.

A simple Monte Carlo EM algorithm can be formulated by rewriting $Q_n(\theta, \theta^{(t-1)})$ as a Monte Carlo sum, i.e.,

$$\begin{aligned} Q_{nm}(\theta, \theta^{(t-1)}) &= \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \log p_\theta(y^{(i)}, z^{(i,j)}) \\ z^{(i,j)} &\sim p_{\theta^{(t-1)}}(z|y^{(i)}). \end{aligned} \tag{56}$$

By the Strong Law of Large Numbers, we are assured that $Q_{nm} \xrightarrow{\text{as}} Q_n$ as $m \to \infty$. If $|\mathcal{Y}| < \infty$, as is often the case for discrete random variables, then we have the pleasing result that $Q_{nm} \xrightarrow{\text{as}} Q_n$ as $n \to \infty$. We refer to the algorithm,

$$\begin{aligned} &\text{E-step:} && z^{(ij)} \sim p_{\theta^{(t-1)}}(\cdot|y^{(i)}),\ j = 1 \dots m, i = 1 \dots n \\ &\text{M-step:} && \theta^{(t)} \leftarrow \arg\max_{\theta \in \Theta} Q_{nm}(\theta, \theta^{(t-1)}) \\ &\text{Loop:} && t \leftarrow t + 1. \end{aligned} \tag{57}$$

as the simple MCEM algorithm and denote the returned value (based on some predefined stopping criteria) as $\hat{\theta}_n$.

However appealing this approach may be, it is certainly not without drawbacks.

First, it remains unclear how large m should be to ensure a reasonable level of accuracy. Setting m too large represents unnecessary computational overhead, while a value too small diminishes the accuracy of the estimate $\hat{\theta}_n$. Second, it may still be the case that sampling from $p_\theta(z|y)$, be it through exact means or Markov Chain Monte Carlo algorithms such as Metropolis-Hastings or Gibbs sampling, may be computationally prohibitive. We address these issues in the following section by introducing importance sampling and incorporating it into the above MCEM algorithm.

4.4.2.1 Importance Sampling

Importance sampling is a standard Monte Carlo technique which allows random variates sampled from one distribution to be used as if they were sampled from another distribution. More precisely, importance sampling makes the following approximation for distributions f, g on $(\mathcal{X}, \mathcal{A}, \mu)$ and statistic h,

$$\begin{aligned} \mathsf{E}_f[h(X)] &= \int_{\mathcal{X}} h(x) f(x) \, \mathrm{d}\mu(x) \\ &\approx \frac{1}{m} \sum_{i=1}^{m} \frac{f(x^{(i)})}{g(x^{(i)})} h(x^{(i)}) \end{aligned} \tag{58}$$

where $\{x^{(i)}\}_1^m$ is an iid sequence with $x^{(1)} \sim g$. From the Strong Law of Large Numbers, it should be clear that the approximation is exact in the limit of large m, provided that $g(x) > 0$ when $f(x) > 0$. Aside from the issue of the support of g, (58) holds for any f and as such, the sequence $\{x^{(i)}\}_1^m$ need not be regenerated for a different f.

Although g can be any distribution with adequate support, it is reasonable to expect some choices of g to be superior to others. Namely, it is desirable to choose a g which ensures f/g is bounded and that the variance of the approximation (58) can be controlled. In general, the selection of such a g can be difficult so we turn to a simple alternative scheme limits the variance of (58) at the expense of increased bias,

viz.,

$$\begin{aligned}\mathsf{E}_f[h(X)] &= \int_{\mathcal{X}} h(x) f(x)\, \mathrm{d}\mu(x) \\ &\approx w_0^{-1} \sum_{i=1}^{m} w_i h(x^{(i)}),\end{aligned} \tag{59}$$

where $w_i = f(x^{(i)})/g(x^{(i)})$ and $w_0 = \sum_i w_i$. By the Strong Law of Large Numbers $w_0/m \xrightarrow{\text{as}} 1$ and (59) remains an asymptotically consistent estimate of $\mathsf{E}_f[h(X)]$. As in (58), we retain the desirable property that the sequence $\{x^{(i)}\}_1^m$ need not be regenerated when f is changed.

In context of the MCEM algorithm (57), the approximation outlined by (59) translates to a one-time generation of a set of latent variates $\{z^{(ij)}\}_{j=1}^m$ for each observation in the observed sequence $\{y^{(i)}\}_1^n$. As such, the added cost of the E-step is avoided and (57) is simplifies to the following Importance Sampled MCEM algorithm,

$$\begin{aligned}&\text{M-step:} && \theta^{(t)} \leftarrow \arg\max_{\theta\in\Theta} Q_{nm}(\theta, \theta^{(t-1)}) \\ &\text{Loop:} && t \leftarrow t+1.\end{aligned} \tag{60}$$

where,

$$Q_{nm}(\theta, \theta^{(t-1)}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{w_{ij}}{w_i} \log p_\theta(z^{(ij)}, y^{(i)}) \tag{61}$$

and,

$$\begin{aligned}w_{ij} &= p_{\theta^{(t-1)}}(z^{(ij)}|y^{(i)}) / p_{\theta^{(0)}}(z^{(ij)}|y^{(i)}) \\ w_i &= \sum_{j=1}^{m} w_{ij} \\ z^{(ij)} &\sim p_{\theta^{(0)}}(\cdot|y^{(i)}),\ i = 1\ldots n,\ j = 1\ldots m.\end{aligned} \tag{62}$$

Henceforth we refer to the procedure characterized by (60) and (61) simply as the MCEM algorithm.

As fortuitous coincidence, we note that the w_{ij}/w_i terms are computationally tractable even when $p_\theta(z|y)$ is not. This is clear from the fact that $p_\theta(z|y)$ is an MRF and the problematic normalization terms cancel, i.e.,

$$\begin{aligned}\frac{w_{ij}}{w_i} &= \frac{p_{\theta^{(t-1)}}(z^{(ij)}|y^{(i)})/p_{\theta^{(0)}}(z^{(ij)}|y^{(i)})}{\sum_{j=1}^{m} p_{\theta^{(t-1)}}(z^{(ij)}|y^{(i)})/p_{\theta^{(0)}}(z^{(ij)}|y^{(i)})} \\ &= \frac{\exp\left\{E_\theta\left(z^{(ij)}, y^{(i)}\right) - E_{\theta^{(0)}}\left(z^{(ij)}, y^{(i)}\right)\right\}}{\sum_{j=1}^{m} \exp\left\{E_\theta\left(z^{(ij)}, y^{(i)}\right) - E_{\theta^{(0)}}\left(z^{(ij)}, y^{(i)}\right)\right\}}.\end{aligned}$$

We summarize by stating that the importance sampled MCEM (60) is a statistically justifiable algorithm for efficiently finding $\hat{\theta}_n$ in latent variable models with a tractable M-step. Fundamentally, the algorithms success hinges upon the coordinate-wise nature of EM coupled with the computational simplicity of importance sampling techniques.

4.4.3 The MCEM+SCL Hybrid

As described above, the Monte Carlo Expectation Maximization (MCEM) algorithm is a statistically motivated procedure for learning HMRF parameters by approximating the E-step of the EM algorithm. Since it is a Monte Carlo based approximation, the technique permits arbitrarily small error through increased sampling. Computationally speaking, the success application of the MCEM algorithm hinges upon the assumptions that a relatively small m permits reasonable accuracy and that the M-step can be efficiently resolved.

On the other hand, the Stochastic Composite Likelihood (SCL) framework, while offering a statistically principled methodology, only simplifies the the M-step, or more generally, eliminates the intractable normalization term from the observed-data MRF.

Our detailed survey of MCEM and SCL informally justifies the following straightforward MCEM+SCL hybrid algorithm,

$$\begin{aligned}&\text{M-step:} && \theta^{(t)} \leftarrow \arg\max_{\theta\in\Theta} \widetilde{Q}_{nm}(\theta, \theta^{(t-1)}) \\ &\text{Loop:} && t \leftarrow t+1.\end{aligned} \tag{63}$$

where,

$$\widetilde{Q}_{nm}(\theta, \theta^{(t-1)}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{c} \frac{\beta_k w_{ij}}{w_i} z_k^{(ij)} \log p_\theta(x_{A_k}^{(ij)} | x_{B_k}^{(ij)})$$

and,

$$\begin{aligned} w_{ij} &= p_{\theta^{(t-1)}}(z^{(ij)}|y^{(i)}) / p_{\theta^{(0)}}(z^{(ij)}|y^{(i)}) \\ w_i &= \sum_{j=1}^{m} w_{ij} \\ x^{(ij)} &= (z^{(ij)}, y^{(i)}) \\ z^{(ij)} &\sim p_{\theta^{(0)}}(\cdot|y^{(i)}). \end{aligned}$$

The MCEM+SCL algorithm can essentially be understood as (60) with the SCL objective rather than the traditional log-likelihood M-step. We use the notation $\widetilde{Q}$ to indicate the SCL surrogate and Q to indicate the traditional EM surrogate.

4.5 Convergence

That the SCL+MCEM hybrid (63) is a viable algorithm for finding SCL-like estimators follows from two convergence arguments. We first demonstrate that the sequence of maximization problems converges to a fixed point, under reasonable conditions. We make this argument through the same techniques which justify EM and MCEM; we reviewing these approaches and show that the hybrid satisfies similar conditions. Next we briefly touch on the statistical convergence of the MCEM+SCL algorithm as a special-case of the SME.

4.5.1 Algorithmic Convergence

In [57], it was first demonstrated that under standard regularity conditions, any iteration of the EM algorithm can be represented as a point-to-point map $M : \Theta \mapsto \Theta$,

i.e.,

$$M_n(\theta^{(t-1)}) = \arg\max\{Q_n(\theta, \theta^{(t-1)}) : \theta \in \Theta\},$$

and that under suitable regularity conditions, the sequence $\{\theta^{(t)} = M_n(\theta^{(t-1)})\}_{t=1}^{\infty}$ converges to the set of stationary points of the log-likelihood, i.e., $\{\theta \in \Theta : \nabla_\theta \ell_n(\theta)\}$.

In [25], the authors characterize EM as discrete-time dynamical system, i.e., $\theta^{(t+1)} = M_n(\theta^{(t)})$. In context, of m-estimation, we could perhaps imagine this as a type of "dynamical m-estimator." Again taking this dynamical system perspective, the authors cast MCEM is a perturbation of the EM system, i.e., $\theta^{(t+1)} = M_n(\theta^{(t)}) + o_{\mathsf{as}}(1)$. As in previous chapters, we use the stochastic order notion $o_{\mathsf{as}}(\cdot)$ to indicate that for a sequence of random variables R_n, the statement $X_n = o_p(R_n)$ means $X_n = Y_n R_n$ for $Y_n \xrightarrow{\text{as}} 0$.

We state a weakened version of the convergence proof of [25, 12] and show how the MCEM+SCL surrogate $\widetilde{Q}_{nm}$ also satisfies the necessary conditions. To this end, it useful to formalize our notion of the ascent function (54) by introducing Lyapunov functions.

Definition 9. Let $\Theta_0 = \{\theta \in \Theta : \theta = M(\theta)\}$ be the set of fixed points of a function $M : \Theta \mapsto \Theta$ (perhaps identical to M_n above). A function $g : \Theta \mapsto \mathbb{R}$ is said to be a *Lyapunov function* relative to (M, Θ) if g is continuous and $g \circ M(\theta) \geq g(\theta)$ for all $\theta \in \Theta$, with equality if and only if $\theta \in \Theta_0$.

In other words, the function M is the basis for an ascent algorithm which maximizes the function g. From the Gibbs Inequality implications of (52), it is clear that the log-likelihood ℓ_n is a Lyapunov function relative to the (combined) M- and E-

steps of (53). That the $sc\ell_n$ objective is Lyapunov follows from,

$$\begin{aligned} sc\ell_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \log \int_{\{y^{(i)}\}\times\mathcal{Z}} \prod_{j=1}^{k} p_\theta\big(x_{A_j}|x_{B_j}\big)^{\beta_j \tilde{z}_j^{(i)}} \,\mathrm{d}\mu(x) \\ &\geq \frac{1}{n}\sum_{i=1}^{n} \int_{\{y^{(i)}\}\times\mathcal{Z}} q(z|y) \log \frac{\prod_{j=1}^{k} p_\theta\big(x_{A_j}|x_{B_j}\big)^{\beta_j \tilde{z}_j^{(i)}}}{q(z|y)} \,\mathrm{d}\mu(x) \\ &= \mathrm{D}_{\mathsf{KL}}\left(p_n(Y)q(Z|Y) \,\middle\|\, \prod_{j=1}^{k} p_\theta(X_{A_j}|X_{B_j})^{\beta_j \tilde{Z}_j} \right) \end{aligned} \tag{64}$$

where as usual $X = (Y, Z)$ and $\tilde{z}^{(i)} \overset{\text{iid}}{\sim} f_\lambda$ is the SCL "selection" random vector. Equation 64 is tight if and only if $q(z|y,\tilde{z}) \propto \prod_{j=1}^{k} p_\theta(X_{A_j}|X_{B_j})^{\beta_j \tilde{z}_j}$ by Gibbs's inequality.[2] Hence, $sc\ell_n$ is Lyapunov with respect to $(\widetilde{M}_n, \Theta)$ where,

$$\widetilde{M}_n(\theta^{(t-1)}) = \arg\max\{\widetilde{Q}_n(\theta, \theta^{(t-1)}) : \theta \in \Theta\}, \tag{65}$$

$$\widetilde{Q}_n(\theta, \theta^{(t-1)}) = \mathrm{D}_{\mathsf{KL}}\Big(p_n(Y) q_{\theta^{(t-1)}}(Z|Y) \,\Big\|\, \textstyle\prod_{j=1}^{k} p_\theta(X_{A_j}|X_{B_j})^{\beta_j \tilde{Z}_j} \Big), \tag{66}$$

and $q_{\theta^{(t-1)}}(Z|Y,\tilde{Z}) \propto \prod_{j=1}^{k} p_{\theta^{(t-1)}}(X_{A_j}|X_{B_j})^{\beta_j \tilde{Z}_j}$.

At a high-level, the proof of the convergence of MCEM [25] can be seen as an extension of the EM convergence proof [57]. Notably the MCEM proof exploits the requirement that the sampling noise in the point-to-point maps tends to zero.

We now state the EM convergence theorem similar to [57] which will serve as a stepping stone for the convergence of the MCEM+SCL hybrid.

Theorem 7. *Let Θ be an open subset of $\mathbb{R}^r$ and let $M_n : \Theta \mapsto \Theta$ be a continuous map with the set $\Theta_0 = \{\theta \in \Theta : \theta = M_n(\theta)\}$ of fixed points. Assume that there exists a Lyapunov function g relative to (M_n, Θ) such that $\{g(\theta) : \theta \in \Theta_0\}$ is a finite set. Let $\mathcal{K} \subset \Theta$ be compact and $\{\theta^{(t)}\}$ a $\mathcal{K}$-valued sequence satisfying,*

$$\lim_{t\to\infty} \left| g(\theta^{(t+1)}) - g \circ M_n(\theta^{(t)}) \right| = 0.$$

[2] The fact that the right-hand side is not a probability is not problematic; the generalized KL-divergence , $\mathrm{D}_{\mathsf{KL}}(P \,\|\, Q) = \int \log \frac{P}{Q} \,\mathrm{d}P - \int \mathrm{d}P + \int \mathrm{d}Q$ is non-negative for all P, Q and zero if and only if $P \equiv Q$. [19]

Then the set $\Theta_0 \cap \mathcal{K}$ *is non-empty, the sequence* $\{g(\theta^{(t)})\}$ *converges to a point* $g_0 \in \{g(\theta) : \theta \in \Theta_0 \cap \mathcal{K}\}$*, and the sequence* $\{\theta^{(t)}\}$ *converges to the set* $[\theta_0] = \{\theta \in \Theta_0 \cap \mathcal{K} : \theta_0 = M_n(\theta)\}$.

Proof. The proof is due to [57] and follows from direct application of Zangwill's Global Convergence Theorem (see Theorem 11.2.3 [12]). □

The convergence of the EM+SCL hybrid characterized by (65) and (66) is proved by Theorem 7 assuming the conditions are satisfied for $\widetilde{M}_n$ and $sc\ell_n$.

We now state and prove the convergence of the SCL+MCEM hybrid as a corollary of Thm. 7. For simplicity of notation we treat the number of training samples as $n = 1$ and omit it from the subscripts. Write $L(\theta) = sc\ell_{n=1}(\theta)$, and,

$$\widetilde{Q}(\theta, \theta^{(t-1)}) = \sum_{j=1}^{k} \tilde{z}_j \beta_j \, \mathsf{E}_{\theta^{(t-1)}}[\log p_\theta(X_{A_j}|X_{B_j})|y],$$

$$\widetilde{Q}_m(\theta, \theta^{(t-1)}) = \sum_{j=1}^{k} \tilde{z}_j \beta_j \, \mathsf{E}_{m,\theta^{(t-1)}}[\log p_\theta(X_{A_j}|X_{B_j})|y],$$

where $\mathsf{E}_{m,\theta}$ denotes the Monte Carlo average of m samples from p_θ; $\mathsf{E}_{m,\theta}$ includes the importance weights. The function $\widetilde{Q}$ is the EM+SCL surrogate and $\widetilde{Q}_m$ is the MCEM+SCL surrogate. Also write,

$$\theta^{(t)} = \widetilde{M}(\theta^{(t-1)}) = \arg\max\{\widetilde{Q}(\theta, \theta^{(t-1)}) : \theta \in \Theta\},$$

$$\theta_m^{(t)} = \widetilde{M}_m(\theta_m^{(t-1)}) = \arg\max\{\widetilde{Q}_m(\theta, \theta_m^{(t-1)}) : \theta \in \Theta\}.$$

The random variable $\theta^{(t)}$ is the EM+SCL estimator and $\theta_m^{(t)}$ is the MCEM+SCL estimator.

Corollary 2. *Let* Θ *be an open subset of* $\mathbb{R}^r$ *and let* $\{p_\theta : \theta \in \Theta\}$ *be a MRF family defined on* m*-dimensional measure space* $(\mathcal{X}, \mathcal{A})$ *with* σ*-finite measure* μ*. Let* $\tilde{z} \sim f_\lambda(\tilde{Z})$ *be the length-k SCL binary selection random vector corresponding to the sequence of* m*-pairs* $\{(A_j, B_j) : \beta_j, f_{\lambda,j} > 0\}$ *and write the observed and latent data as* $X = (Y, Z)$*.*

Additionally, make the following assumptions.

A1 *The map* $(\theta, \theta') \mapsto \mathsf{E}_{\theta'}[\log p_\theta(X_{A_j}|X_{B_j})|y]$ *is finite and continuous on* Θ^2 *for all* $j = 1 \dots k$.

A2 *The set* $\{L(\theta) : \theta \in \Theta_0\}$ *is finite, where* $\Theta_0 = \{\theta \in \Theta : \theta = \widetilde{M}(\theta)\}$, *and the closure of the sequence* $\{\theta^{(t)}\}$ *is a compact subset of* Θ.

A3 *The convergence of the Monte Carlo composition is uniformly strong, i.e., for* $d \to \infty$, $\sup_{\theta \in \Theta} \left| L \circ \widetilde{M}_d(\theta) - L \circ \widetilde{M}(\theta) \right| = o_{as}(1)$.

A4 *The sequence of m-pairs* $\{(A_j, B_j) : \lambda_j, \beta_j > 0\}$ *ensures the identifiability of* $p_\theta(X)$. *Additionally, the selection random variable is* $|\tilde{z}| < \infty$.

Then, the sequence $\{L(\theta^{(t)})\}_t$ *has a limit and, the sequence* $\{\theta^{(t)}\}$ *converges to a member of the set* Θ_0, *almost surely.*

Proof. From A1, A2, and (64), it is clear that L is Lyapunov relative to $(\widetilde{M}, \Theta)$. Therefore, each iteration of this SCL+EM algorithm increases the objective, $L \circ \widetilde{M}(\theta) \geq L(\theta)$, with equality if and only if $\theta \in \Theta_0$. Because $\widetilde{M}$ is continuous and the Monte Carlo composition converges uniformly (A3), the proof follows from Theorem 7. That the stationary point is the MLE in the case follows □

We note that this theorem indeed generalizes the proofs of convergence for EM and MCEM through particular choices of f_λ and β, e.g., $\beta_{\mathsf{FL}} = f_\lambda(z_{\mathsf{FL}}) = 1$.

Obviously Assumption 3 is fairly strong. It appears that [25] is the most general treatment of the subject. The authors relax Assumption 3, by restricting the proof of MCEM convergence to curved exponential families. They show that, even when M_d is computed from ergodic rather than independent samples, assumption 3 is satisfied. Restricting attention to exponential families and M_d from a Monte Carlo average rather than an MCMC average, reveals the key properties which eliminate A3. In context of, MCEM+SCL the A3 condition is met when the number of Monte Carlo samples is increased with each iteration.

4.5.2 Statistical Convergence

In the previous section, it was proved that the EM+SCL hybrid algorithm converges to a maximum when the number Monte Carlo samples are increased after each iteration. We now ask whether or not the M-step is capable of recovering the population maximizer, that is, we address the (statistical) consistency of the EM+SCL M-step. That the M-Step is consistent, follows trivially from the fact that the limit point of the previous proof is simply the SCL objective on the marginal $p_\theta(Y)$.

4.6 Gradient-Based Alternative

Let $X = (Y, Z)$ and $J = \{j \in \mathbb{Z}_+ : \beta_j > 0, f_\lambda(\tilde{Z}_j) > 0\}$. Suppose that $|J| < \infty$ and each member of the set $\{p_\theta(X_{A_j}|X_{B_j})\}_{j\in J}$ of conditionals is positive and differentiable in θ. Furthermore, assume there exists a function $K(x,\theta)$ such that $\int |K(x,\theta)|\, d\mu(z) < \infty$ and $\left|\nabla_\theta p_\theta(x_{A_j}|x_{B_j})\right|_{\theta=\theta'} \leq K(x,\theta)$ for all θ' in a neighborhood of θ. Finally, assume that the SCL selection random vector and corresponding weights are of bounded length, i.e., $\sup_{\tilde{Z}} \|\tilde{Z}\| < \infty$ and $\sup_\beta \|\beta\| < \infty$.

These assumptions satisfy the conditions of the Lebesgue dominated convergence theorem applied to the SCL objective on the marginal $p_\theta(y)$. Thus, we may conclude,

$$
\begin{aligned}
&\nabla_\theta \log \int_{y\times\mathcal{Z}} \prod_{j=1}^k p_\theta(x_{A_j}|x_{B_j})^{\beta_j\tilde{z}_j}\, \mathrm{d}\mu(x) = \\
&= \frac{1}{\int_{\mathcal{X}} \prod_{j=1}^k p_\theta(x_{A_j}|x_{B_j})^{\beta_j\tilde{z}_j}\, \mathrm{d}\mu(x)} \nabla_\theta \int_{y\times\mathcal{Z}} \prod_{j=1}^k p_\theta(x_{A_j}|x_{B_j})^{\beta_j\tilde{z}_j}\, \mathrm{d}\mu(x) \\
&= \int_{y\times\mathcal{Z}} \frac{\prod_{j=1}^k p_\theta(x_{A_j}|x_{B_j})^{\beta_j\tilde{z}_j}}{\int_{\mathcal{X}} \prod_{j=1}^k p_\theta(x_{A_j}|x_{B_j})^{\beta_j\tilde{z}_j}\, \mathrm{d}\mu(x)} \nabla_\theta \log \prod_{j=1}^k p(x_{A_j}|x_{B_j})^{\beta_j\tilde{z}_j}\, \mathrm{d}\mu(x) \\
&= \sum_{j=1}^k \beta_j\tilde{z}_j \int_{y\times\mathcal{Z}} q_\theta(z|y) \nabla_\theta \log p(x_{A_j}|x_{B_j})\, \mathrm{d}\mu(x) \\
&= \sum_{j=1}^k \beta_j\tilde{z}_j\, \mathsf{E}_{q_\theta(z|y)} \left[\nabla_\theta \log p(x_{A_j}|x_{B_j})\middle| y\right] \qquad (67)
\end{aligned}
$$

where $q_\theta(Z|Y) \propto \prod_{j=1}^k p_\theta(x_{A_j}|x_{B_j})^{\beta_j\tilde{z}_j}$.

Much like the preceding sections one may approximate (67) using a Monte Carlo average with samples drawn from $q_\theta(Z|Y)$. It is also natural to apply the derivation of (67) to develop a stochastic gradient descent algorithm. For example, consider the Robbins-Monro procedure [42],

$$\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} + \gamma_t \nabla_\theta sc\ell_1(y^{(t)}, z^{(t)}; \theta^{(t-1)}) \tag{68}$$

where $z^{(t)} \sim q_\theta(Z|y^{(t)})$ is importance sampled in a fashion similar to the preceding section. Here γ_t is the learning rate and must be suitable chosen to ensure convergence. For more details see [12].

4.7 Empirical Study

In this section we validate the theoretical analysis related to the MCEM+SCL algorithm outlined by (63). We evaluate the performance of two SCL policies in learning the parameters of a latent linear-chain MRF. These models are often referred to as "hidden Markov models" but due to the ambiguity of this term, we prefer the term latent linear-chain MRF or latent Boltzmann chain (see Section 3.8.3.1 for more details).

Recall that a length T fully observed Boltzmann chain (BC) is given by,

$$\begin{aligned} p_\theta(z, y) &= p_\theta(z_1)p_\theta(y_1|z_1)\prod_{t=2}^{T} p_\theta(z_t|z_{t-1})p_\theta(y_t|z_t) \\ &= \pi(z_1)B(z_1, y_1)\prod_{t=2}^{T} A(z_{t-1}, z_t)B(z_t, y_t). \end{aligned}$$

Here, the random variate z corresponds to the sequence of transition states while the random variate y corresponds to the sequence of emission states. Denoting the number of transition states as s and the number of emission states as m, the parameters $\theta = (\pi, A, B)$ correspond to the distribution of the start state ($s \times 1$), transition matrix ($s \times s$), and emission matrix ($s \times m$). The latent Boltzmann chain (LBC) is simply the BC with transition states marginalized out, i.e., $p_\theta(y) = \sum_{z \in \mathcal{Z}^T} p_\theta(y, z)$.

As in Chapter 3, this is a convenient model for validation as efficient algorithms exist for computing the full normalization term.[12] For example, the Baum-Welch algorithm has a complexity of $O(s^2T)$; naive computation has a complexity exponential in chain length, i.e., $O(s^TT)$.[23]

For this experiment we used a randomly parameterized, $T = 20$ length chain with $s = 10$ transition states and $m = 100$ emission states. We generated $n_0 = 1000$ training samples and $n_1 = 5000$ testing examples using a Gibbs sampler. We used a fixed weight $\beta_j = 1$ for all components; we expect results to improve using the heuristic presented in Chapter 3.

The left column of Figure 23 depicts the training and testing negative log-likelihood resulting from varying the probability λ_{FL} of selecting the full-likelihood component during M-step SCL policy. Likewise, the right column of Figure 23 depicts the effect of varying the probability λ_{PL1} of computing the order-2 pseudo-likelihood component independently at each site $t = 1 \dots 20$. In both cases the order-1 pseudo-likelihood was always computed, i.e., $\lambda_{PL1} = 1$. Both experiments employed the importance weighting scheme outlined above; the number of MCEM samples was increased by 1, every 5 iterations. The results are averaged over 10 runs; all results are statistically significant by the 25th iteration.

At a high-level, we observe trends quite similar to those of Chapter 3. These results are noticeably less smooth–a consequence of the Monte Carlo nature of the E-step. In both the PL1/PL2 and PL1/FL policies the test set negative log-likelihood stabilizes above the train set negative log-likelihood (as expected) and in order of increasing computation (also as expected). Also noticeable, the randomness diminishes with the number of iterations as the number of importance samples is increased.

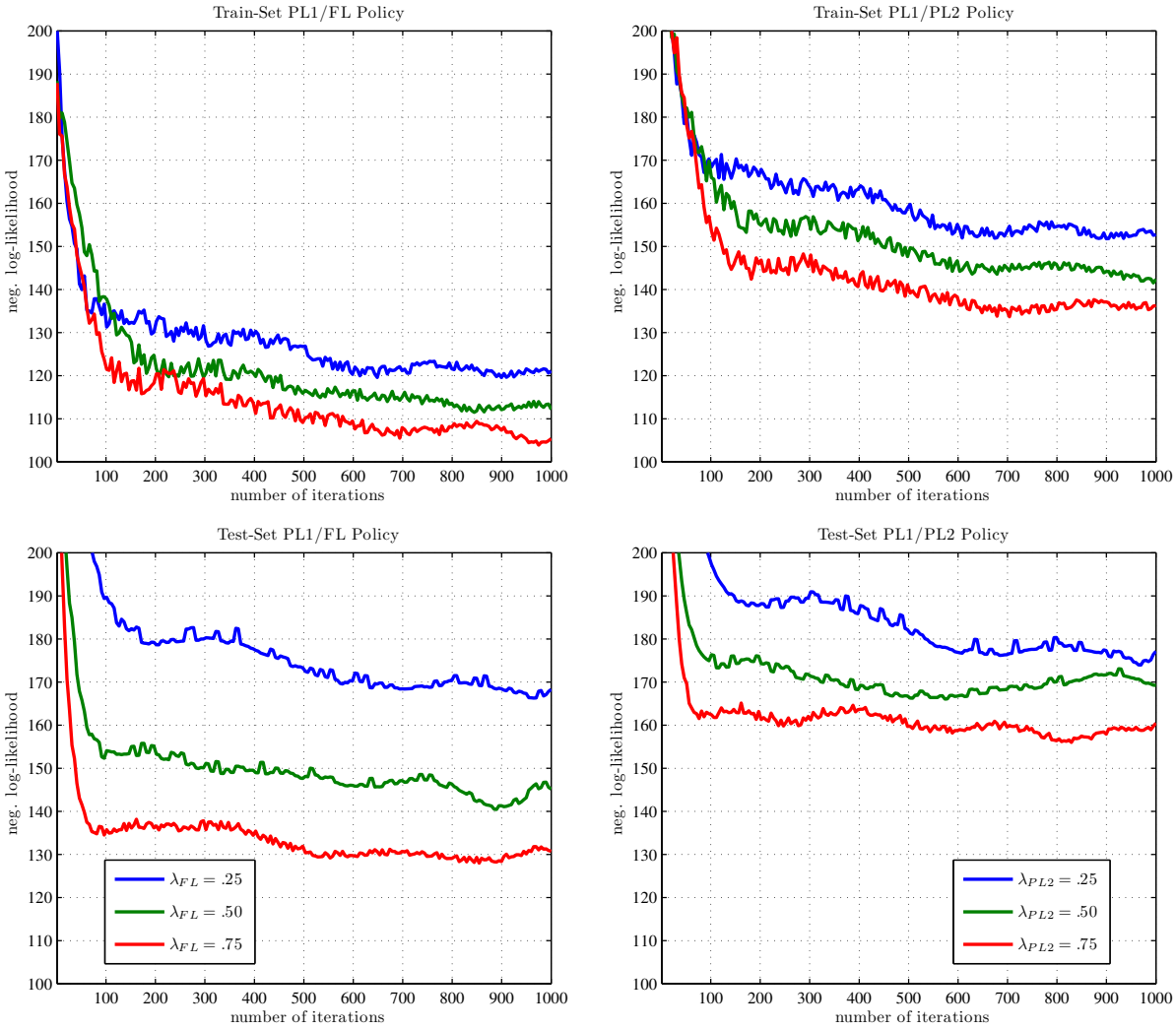

Figure 23: Train (top row) and test (bottom row) negative log-likelihood (y-axis) after a given number of iterations (x-axis) of the MCEM+SCL hybrid algorithm (63). See Section 4.7 for discussion.

CHAPTER V

LABELING COST: GENERATIVE SEMI-SUPERVISED LEARNING

5.1 Introduction

Semi-supervised learning (SSL) is a technique for estimating statistical models using both labeled and unlabeled data. It is particularly useful when the costs of obtaining labeled and unlabeled samples are different. In particular, assuming that unlabeled data is more easily available, SSL provides improved modeling accuracy by adding a large number of unlabeled samples to a relatively small labeled dataset.

The practical value of SSL has motivated several attempts to mathematically quantify its value beyond traditional supervised techniques. Of particular importance is the dependency of that improvement on the amount of unlabeled and labeled data. In the case of structured prediction the accuracy of the SSL estimator depends also on the specific manner in which sequences are labeled. Focusing on the framework of generative or likelihood-based SSL applied to classification and structured prediction we identify the following questions which we address in this chapter.

Q1: *Consistency (classification).* What combinations of labeled and unlabeled data lead to precise models in the limit of large data.

Q2: *Accuracy (classification).* How can we quantitatively express the estimation accuracy for a particular generative model as a function of the amount of labeled and unlabeled data. What is the improvement in estimation accuracy resulting from replacing an unlabeled example with a labeled one.

Q3: *Consistency (structured prediction).* What strategies for sequence labeling lead

to precise models in the limit of large data.

Q4: *Accuracy (structured prediction).* How can we quantitatively express the estimation quality for a particular model and structured labeling strategy. What is the improvement in estimation accuracy resulting from replacing one labeling strategy with another.

Q5: *Tradeoff (classification and structured prediction).* How can we quantitatively express the tradeoff between the two competing goals of improved prediction accuracy and low labeling cost. What are the possible ways to resolve that tradeoff optimally within a problem-specific context.

Q6: *Practical Algorithms.* How can we determine how much data to label in practical settings.

The first five questions are of fundamental importance to SSL theory. Recent related work has concentrated on large deviation bounds for discriminative SSL as a response to Q1 and Q2 above. While enjoying broad applicability, such non-parametric bounds are weakened when the model family's worst-case is atypical. By forgoing finite sample analysis, our approach complements these efforts and provides insights which apply to the specific generative models under consideration. In presenting answers to the last question, we reveal the relative merits of asymptotic analysis and how its employ, perhaps surprisingly, renders practical heuristics for controlling labeling cost.

Our asymptotic derivations are possible by extending the recently proposed stochastic composite likelihood formalism [21] and showing that generative SSL is a special case of that extension. The implications of this analysis are demonstrated using a simulation study as well as text classification and NLP structured prediction experiments. The developed framework, however, is general enough to apply to any

generative SSL problem. As in [35], the delta method transforms our results from parameter asymptotics to prediction risk asymptotics.

5.2 Related Work

Semi-supervised learning has received much attention in the past decade. Perhaps the first study in this area was done by [14] who examined the convergence of the classification error rate as a labeled example is added to an unlabeled dataset drawn from a Gaussian mixture model. [40] proposed a practical SSL framework based on maximizing the likelihood of the observed data. An edited volume describing more recent developments is [15].

The goal of theoretically quantifying the effect of SSL has recently gained increased attention. In [16], the authors use asymptotic bias arguments to analyze generative SSL under various scenarios of labeled and unlabeled data while [59] considers the asymptotic efficiency. [47] examined the effect of using unlabeled samples with imperfect models for mixture models. [4] and [46] analyze discriminative SSL using PAC theory and large deviation bounds. Additional analysis has been conducted under specific distributional assumptions such as the "cluster assumption", "smoothness assumption" and the "low density assumption."[15] However, many of these assumptions are criticized by [5].

Excluding the works of Cohen & Cozman and Zhang, this chapter complements the above studies by focusing on generative rather than discriminative SSL. Whereas Cohen & Cozman consider the scenario in which error is dominated by bias rather than variance, we analyze and empirically motivate the reverse. This chapter generalizes that of Zhang's by including SSL for structured prediction tasks. Moreover, this leads to heuristics for optimally obtaining partially labeled samples. In contrast to most other studies, we derive model specific asymptotics as opposed to non-parametric large deviation bounds. While such bounds are helpful as they apply to a broad set

of cases, they also provide less information than model-based analysis due to their generality. Our analysis, on the other hand, requires knowledge of the specific model family and an estimate of the model parameters. The resulting asymptotics, apply specifically to the case at hand without the need of potentially loose bounds.

This chapter will explore and answer questions Q1-Q6 in the context of generative SSL. In particular, it provides a new framework for examining the accuracy-cost SSL tradeoff in a way that is quantitative, practical, and model-specific.

5.3 Stochastic SSL Estimators

Generative SSL [40, 15] estimates a parametric model by maximizing the observed likelihood incorporating L labeled and U unlabeled examples

$$\ell(\theta) = \sum_{i=1}^{L} \log p_\theta(X^{(i)}, Y^{(i)}) + \sum_{i=L+1}^{L+U} \log p_\theta(X^{(i)}) \tag{69}$$

where $p_\theta(X^{(i)})$ above is obtained by marginalizing the latent label $\sum_y p_\theta(X^{(i)}, y)$. A classical example is the naive Bayes model in [40] where $p_\theta(X, Y) = p_\theta(X|Y)p(Y)$, $p_\theta(X|Y = y) = \text{Mult}([\theta_y]_1, \ldots, [\theta_y]_V)$. The framework, however, is general enough to apply to any generative model $p_\theta(X, Y)$.

To analyze the asymptotic behavior of the maximizer of (69) we assume that the ratio between labeled to unlabeled examples $\lambda = L/(L + U)$ is kept constant while $n = L + U \to \infty$. More generally, we assume a stochastic version of (69) where each one of the n samples $X^{(1)}, \ldots, X^{(n)}$ is labeled with probability λ

$$\begin{aligned} \ell_n(\theta) = &\sum_{i=1}^{n} Z^{(i)} \log p_\theta(X^{(i)}, Y^{(i)}) \\ &+ \sum_{i=1}^{n} (1 - Z^{(i)}) \log p_\theta(X^{(i)}), \quad Z^{(i)} \sim \text{Bin}(1, \lambda). \end{aligned} \tag{70}$$

The variable $Z^{(i)}$ above is an indicator taking the value 1 with probability λ and 0 otherwise. Due to the law of large numbers for large n we will have approximately $L = n\lambda$ labeled samples and $U = n(1 - \lambda)$ unlabeled samples thus achieving the asymptotic behavior of (69).

Equation (70) is sufficient to handle the case of classification. However, in the case of structured prediction we may have sequences $X^{(i)}, Y^{(i)}$ where for each i some components of the label sequence $Y^{(i)}$ are missing and some are observed. For example one label sequence may be completely observed, another may be completely unobserved, and a third may have the first half labeled and the second half not.

More formally, we assume the existence of a sequence labeling policy or strategy $\wp$ which maps label sequences $Y^{(i)} = (Y_1^{(i)}, \ldots, Y_m^{(i)})$ to a subset corresponding to the observed labels $\wp(Y^{(i)}) \subset \{Y_1^{(i)}, \ldots, Y_m^{(i)}\}$. To achieve generality we allow the labeling policy $\wp$ to be stochastic, leading to different subsets of $\{Y_1^{(i)}, \ldots, Y_m^{(i)}\}$ with different probabilities. A simple "all or nothing" labeling policy could label the entire sequence with probability λ and otherwise ignore it. Another policy may label the entire sequence, the first half, or ignore it completely with equal probabilities

$$\wp(Y) = \begin{cases} Y_1^{(i)}, \ldots, Y_m^{(i)} & \text{with probability } 1/3 \\ \emptyset & \text{with probability } 1/3 \\ Y_1^{(i)}, \ldots, Y_{\lfloor m/2 \rfloor}^{(i)} & \text{with probability } 1/3 \end{cases} . \tag{71}$$

We thus have the following generalization of (70) for structured prediction

$$\ell_n(\theta) = \sum_{i=1}^{n} \log p_\theta(\wp(Y^{(i)}), X^{(i)}). \tag{72}$$

Equation (72) generalizes standard SSL from all or nothing labeling to arbitrary labeling policies. The fundamental SSL question in this case is not simply what is the dependency of the estimation accuracy on n and λ. Rather we ask what is the dependency of the estimation accuracy on the labeling policy $\wp$. Of particular interest is the question what labeling policies $\wp$ achieve high estimation accuracy coupled with low labeling cost. Answering these questions leads to a generative SSL theory that quantitatively balances estimation accuracy and labeling cost.

Finally, we note that both (70) and (72) are random variables whose outcomes

depend on the random variables $Z^{(1)}, \ldots, Z^{(n)}$ (for (70)) or $\wp$ (for (72)). As a consequence, the analysis of the maximizer $\hat{\theta}_n$ of (70) or (72) needs to be done in a probabilistic manner.

5.4 A1: Consistency (Classification)

Assuming that the data is generated from $p_{\theta_0}(X, Y)$, consistency corresponds to the convergence of

$$\hat{\theta}_n = \arg\max_{\theta} \ell_n(\theta) \tag{73}$$

to θ_0 with probability 1 as $n \to \infty$ (ℓ_n is defined in (70)). This implies that in the limit of large data our estimator would converge to the truth. Note that large data $n \to \infty$ in this case means that both labeled and unlabeled data increase to ∞ (but their relative sizes remain the constant λ).

We show in this section that the maximizer of (70) is consistent assuming that $\lambda > 0$. This is not an unexpected conclusion but for the sake of completeness we prove it rigorously. The proof technique is also used when we discuss consistency of SSL estimators for structured prediction. The notation $\xrightarrow{\text{p}}$ denotes convergences in probability.[24]

Definition 10. A distribution $p_\theta(X, Y)$ is said to be identifiable if $\theta \neq \eta$ entails that $p_\theta(X, Y) - p_\eta(X, Y)$ is not identically zero.

Proposition 7. *Let $\Theta \subset \mathbb{R}^r$ be a compact set, and $p_\theta(x, y) > 0$ be identifiable and smooth in θ. Then if $\lambda > 0$ the maximizer $\hat{\theta}_n$ of (70) is consistent i.e., $\hat{\theta}_n \to \theta_0$ as $n \to \infty$ with probability 1.*

Proof. The likelihood function, modified slightly by a linear combination with a constant is

$$\ell'_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(Z^{(i)}\log p_\theta(X^{(i)},Y^{(i)}) - \lambda\log p_{\theta_0}(X^{(i)},Y^{(i)})\right) + \frac{1}{n}\sum_{i=1}^{n}\left((1-Z^{(i)})\log p_\theta(X^{(i)}) - (1-\lambda)\log p_{\theta_0}(X^{(i)})\right),$$

which converges by the the strong law of large numbers as $n \to \infty$ to its expectation with probability 1, i.e.,

$$\ell'_n(\theta) \xrightarrow{\text{P}} \mu(\theta) = -\lambda D(p_{\theta_0}(X,Y)||p_\theta(X,Y)) - (1-\lambda)D(p_{\theta_0}(X)||p_\theta(X))).$$

If we restrict ourselves to the compact set $S = \{\theta : c_1 \leq \|\theta - \theta_0\| \leq c_2\}$ then $|\log p_\theta(X,Y)| < K(X,Y) < \infty$, $\forall\theta \in S$. As a result, the conditions for the uniform strong law of large numbers, cf. chapter 16 of [24], hold on S leading to

$$P\left\{\lim_{n\to\infty}\sup_{\theta\in S}|\ell'_n(\theta) - \mu(\theta)| = 0\right\} = 1. \tag{74}$$

Since $p_\theta(X,Y)$ is identifiable, we have $D(p_{\theta_0}(X,Y)||p_\theta(X,Y)) \geq 0$ with equality iff $\theta = \theta_0$. Since also $D(p_{\theta_0}(X)||p_\theta(X))) \geq 0$ we have that $\mu(\theta) \leq 0$ with equality iff $\theta = \theta_0$ (assuming $\lambda > 0$). Furthermore, since the function $\mu(\theta)$ is continuous it attains its negative supremum on the compact S: $\sup_{\theta\in S}\mu(\theta) < 0$.

Combining this fact with (74) we have that there exists N such that for all $n > N$ the likelihood maximizers on S achieves strictly negative values of $\ell'_n(\theta)$ with probability 1. However, since $\ell'_n(\theta)$ can be made to achieve values arbitrarily close to zero under $\theta = \theta_0$, we have that $\hat{\theta}_n \notin S$ for $n > N$. Since c_1, c_2 were chosen arbitrarily $\hat{\theta}_n \to \theta_0$ with probability 1. □

The proof follows the consistency proof of Chapter 3 with the exception that it does not assume independence of the indicator functions $Z^{(i)}$ and $(1 - Z^{(i)})$, as is assumed there.

The above proposition is not surprising. As $n \to \infty$ the number of labeled examples increase to ∞ and thus it remains to ensure that adding an increasing number of unlabeled examples does not hurt the estimator. More interesting is the quantitative description of the accuracy of $\hat{\theta}_n$ and its dependency on θ_0, λ, n which we turn to next.

5.5 A2: Accuracy (Classification)

The proposition below states that the distribution of the maximizer of (70) is asymptotically normal and provides its variance which may be used to characterize the accuracy of $\hat{\theta}_n$ as a function of n, θ_0, λ. As in Section 5.4 our proof proceeds by casting generative SSL as an extension of stochastic composite likelihood.

In Proposition 8 (below) and in Proposition 10 we use $\mathsf{Var}_{\theta_0}(H)$ to denote the variance matrix of a random vector H under p_{θ_0}. The notation $\rightsquigarrow$ denotes convergences in distribution [24] and $\nabla f(\theta)$, $\nabla^2 f(\theta)$ are the $r \times 1$ gradient vector and $r \times r$ matrix of second order derivatives of $f(\theta)$.

Proposition 8. *Under the assumptions of Proposition 7 as well as convexity of* Θ *we have the following convergence in distribution of the maximizer of* (70)

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N\left(0, \Sigma^{-1}\right) \tag{75}$$

as $n \to \infty$*, where*

$$\Sigma = \lambda\, \mathsf{Var}_{\theta_0}(V_1) + (1-\lambda)\, \mathsf{Var}_{\theta_0}(V_2)$$

$$V_1 = \nabla_\theta \log p_{\theta_0}(X, Y), \quad V_2 = \nabla_\theta \log p_{\theta_0}(X).$$

Proof. By the mean value theorem and convexity of Θ, there is $\eta \in (0,1)$ for which $\theta' = \theta_0 + \eta(\hat{\theta}_n - \theta_0)$ and

$$\nabla \ell_n(\hat{\theta}_n) = \nabla \ell_n(\theta_0) + \nabla^2 \ell_n(\theta')(\hat{\theta}_n - \theta_0).$$

Since $\hat{\theta}_n$ maximizes ℓ_n we have $\nabla \ell_n(\hat{\theta}_n) = 0$ and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\sqrt{n}\left(\nabla^2 \ell_n(\theta')\right)^{-1} \left(\nabla \ell_n(\theta_0)\right). \tag{76}$$

By Proposition 7 we have $\hat{\theta}_n \xrightarrow{\text{p}} \theta_0$ which implies that $\theta' \xrightarrow{\text{p}} \theta_0$ as well. Furthermore, by the law of large numbers and the fact that $W_n \xrightarrow{\text{p}} W$ implies $g(W_n) \xrightarrow{\text{p}} g(W)$ for continuous g,

$$\begin{aligned}(\nabla^2 \ell_n(\theta'))^{-1} &\xrightarrow{\text{p}} (\nabla^2 \ell_n(\theta_0))^{-1} \\ &\xrightarrow{\text{p}} \Big(\lambda \,\mathsf{E}_{\theta_0} \nabla^2 \log p_{\theta_0}(X,Y) \\ &\quad + (1-\lambda)\,\mathsf{E}_{\theta_0} \nabla^2 \log p_{\theta_0}(X)\Big)^{-1} = \Sigma^{-1}\end{aligned} \tag{77}$$

where in the last equality we used a well known identity concerning the Fisher information.

For the remaining term in the rhs of (76) we have

$$-\sqrt{n}\nabla \ell_n(\theta_0) = -\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}(W^{(i)} + Q^{(i)}) \tag{78}$$

where $W^{(i)} = Z^{(i)}\nabla \log p_{\theta_0}(X^{(i)}, Y^{(i)})$, $Q^{(i)} = (1 - Z^{(i)})\nabla \log p_{\theta_0}(X^{(i)})$. Since (78) is an average of iid random vectors $W^{(i)} + Q^{(i)}$ it is asymptotically normal by the central limit theorem with mean

$$\begin{aligned}\mathsf{E}_{\theta_0}(Q + W) &= \lambda \,\mathsf{E}_{\theta_0} \nabla \log p_{\theta_0}(X,Y) \\ &\quad + (1-\lambda)\,\mathsf{E}\,\nabla \log p_{\theta_0}(X) = \lambda 0 + (1-\lambda)0.\end{aligned}$$

and variance

$$\begin{aligned}\mathsf{Var}_{\theta_0}(W + Q) &= \mathsf{E}_{\theta_0} W^2 + \mathsf{E}_{\theta_0} Q^2 + 2\,\mathsf{E}_{\theta_0} WQ \\ &= \lambda \,\mathsf{Var}_{\theta_0} V_1 + (1-\lambda)\,\mathsf{Var}_{\theta_0} V_2\end{aligned}$$

where we used the fact that $\mathsf{E}(Z(1-Z)) = \mathsf{E}\,Z - \mathsf{E}\,Z^2 = 0$ for binary random variables Z.

We have thus established that

$$-\sqrt{n}\nabla \ell_n(\theta_0) \rightsquigarrow N(0, \Sigma). \tag{79}$$

We finish the proof by combining (76), (77) and (79) using Slutsky's theorem. □

Proposition 8 characterizes the asymptotic estimation accuracy using the matrix Σ. Two convenient one dimensional summaries of the accuracy are the trace and the determinant of Σ. In some simple cases (such as binary event naive Bayes) $\mathsf{tr}(\Sigma)$ can be brought to a mathematically simple form which exposes its dependency on θ_0, n, λ. In other cases the dependency may be obtained using numerical computing.

Figure 24 displays three error measures for the multinomial naive Bayes SSL classifier [40] and the Reuters RCV1 text classification data. In all three figures the error measures are represented as functions of n (horizontal axis) and λ (vertical axis). The error measures are classification error rate (left), trace of the empirical MSE (middle), and log-trace of the asymptotic variance (right). The measures were obtained over held-out sets and averaged using cross validation. Figure 26 (middle) displays the asymptotic variance as a function of n and λ for a randomly drawn θ_0.

As expected the measures decrease with n and λ in all the figures. It is interesting to note, however, that the shapes of the contour plots are very similar across the three different measures (top row). This confirms that the asymptotic variance (right) is a valid proxy for the finite sample measures of error rates and empirical MSE. We thus conclude that the asymptotic variance is an attractive measure that is similar to finite sample error rate and at the same time has a convenient mathematical expression.

5.6 A3: Consistency (Structured)

In the case of structured prediction the log-likelihood (72) is specified using a stochastic labeling policy. In this section we consider the conditions on that policy that ensure estimation consistency, i.e., convergence of the maximizer of (72) to θ_0 as $n \to \infty$.

We assume that the labeling policy $\wp$ is a probabilistic mixture of deterministic sequence labeling functions $\chi_1, \ldots, \chi_k$. In other words, $\wp(Y)$ takes values $\chi_i(Y), i = 1, \ldots, k$ with probabilities $\lambda_1, \ldots, \lambda_k$. For example the policy (71) corresponds to $\chi_1(Y) = Y$, $\chi_2(Y) = \emptyset$, $\chi_3(Y) = \{Y_1, \ldots, Y_{\lfloor m/2 \rfloor}\}$ (where $Y = \{Y_1, \ldots, Y_m\}$) and

$\lambda = (1/3, 1/3, 1/3)$.

Using the above notation we can write (72) as

$$\ell_n(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{k} Z_j^{(i)} \log p_\theta(\chi_j(Y^{(i)}), X^{(i)}) \tag{80}$$

$$Z^{(i)} \sim \text{Mult}(1, (\lambda_1, \ldots, \lambda_k))$$

which exposes its similarity to the stochastic composite likelihood function in [21]. Note however that (80) is not formally a stochastic composite likelihood since $Z_j^{(i)}, j = 1, \ldots, k$ are not independent and since $\chi_j(Y)$ depends on the length of the sequence Y (see for example χ_1 and χ_3 above). We also use the notation S_j^m for the subset of labels provided by χ_j on length-m sequences

$$\chi_j(Y_1, \ldots, Y_m) = \{Y_i : i \in S_j^m\}.$$

Definition 11. A labeling policy is said to be identifiable if the following map is injective

$$\bigcup_{m:q(m)>0} \bigcup_{j=1}^{k} \{p_\theta(\{Y_r : r \in S_j^m\}, X)\} \to p_\theta(X, Y)$$

where q is the distribution of sequences lengths. In other words, there is at most one collection of probabilities corresponding to the lhs above that does not contradict the joint distribution.

The importance of Definition 11 is that it ensures the recovery of θ_0 from the sequences partially labeled using the labeling policy. For example, a labeling policy characterized by $\chi_1(Y) = Y_1$, $\lambda_1 = 1$ (always label only the first sequence element) is non-identifiable for most interesting p_θ as the first sequence component is unlikely to provide sufficient information to characterize the parameters associated with transitions $Y_t \to Y_{t+1}$.

Proposition 9. *Assuming the conditions of Proposition 7, and* $\lambda_1, \ldots, \lambda_k > 0$ *with identifiable* $\chi_1, \ldots, \chi_k$*, the maximizer of* (80) *is consistent i.e.,* $\hat{\theta}_n \to \theta_0$ *as* $n \to \infty$ *with probability 1.*

Proof. The log-likelihood (72), modified slightly by a linear combination with a constant is

$$\ell'_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \Big(Z_j^{(i)} \log p_\theta(\chi_j(Y^{(i)}), X^{(i)}) - \lambda_j \log p_{\theta_0}(\chi_j(Y^{(i)}), X^{(i)}) \Big).$$

By the strong law of large numbers $\ell'_n(\theta)$ converges to its expectation

$$\mu(\theta) = -\sum_{j=1}^{k} \lambda_j \sum_{m>0} q(m) \cdot D(p_{\theta_0}(\{Y_i : i \in S_j^m\}, X) || p_\theta(\{Y_i : i \in S_j^m\}, X)).$$

Since μ is a linear combination of KL divergences with positive weights it is non-negative and is 0 if $\theta = \theta_0$. The identifiability of the labeling policy ensures that $\mu(\theta) > 0$ if $\theta \neq \theta_0$. We have thus established that $\ell_n(\theta)$ converges to a non-negative continuous function $\mu(\theta)$ whose maximum is achieved at θ_0. □

As with Proposition 7, the proof follows the consistency proof of 3 with the exception that it does not assume independence of the indicator functions $Z^{(i)}$ and $(1 - Z^{(i)})$.

Ultimately, the precise conditions for consistency will depend on the parametric family p_θ under consideration. For many structured prediction models such as Markov random fields the consistency conditions are mild. Depending on the precise feature functions, consistency is generally satisfied for every policy that labels contiguous subsequences with positive probability. However, some care need be applied for models like HMMs which contain parameters associated with start and/or end labels and with models asserting higher order Markov assumptions.

5.7 A4: Accuracy (Structured)

We consider in this section the dependency of the estimation accuracy in structured prediction SSL (72) on n, θ_0 but perhaps most interestingly on the labeling policy $\wp$.

Doing so provides insight into not only how much data to label but also in what way.

Proposition 10. *Under the assumptions of Proposition 9 as well as convexity of* Θ *we have the following convergence in distribution of the maximizer of* (80)

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N\left(0, \Sigma^{-1}\right) \tag{81}$$

as $n \to \infty$*, where*

$$\Sigma = \mathsf{E}_{q(m)} \left\{ \sum_{j=1}^{k} \lambda_j \, \mathsf{Var}_{\theta_0}(\nabla V_{jm}) \right\}$$

$$V_{jm} = \log p_{\theta_0}(\{Y_i : i \in S_j^m\}, X).$$

Proof. By the mean value theorem and convexity of Θ there is $\eta \in (0,1)$ for which $\theta' = \theta_0 + \eta(\hat{\theta}_n - \theta_0)$ and

$$\nabla \ell_n(\hat{\theta}_n) = \nabla \ell_n(\theta_0) + \nabla^2 \ell_n(\theta')(\hat{\theta}_n - \theta_0).$$

Since $\hat{\theta}_n$ maximizes ℓ, $\nabla \ell_n(\hat{\theta}_n) = 0$ and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\sqrt{n}(\nabla^2 \ell_n(\theta'))^{-1} \nabla \ell_n(\theta_0). \tag{82}$$

By Proposition 9 we have $\hat{\theta}_n \xrightarrow{\text{p}} \theta_0$ which implies that $\theta' \xrightarrow{\text{p}} \theta_0$ as well. Furthermore, by the law of large numbers and the fact that if $W_n \xrightarrow{\text{p}} W$ then $g(W_n) \xrightarrow{\text{p}} g(W)$ for continuous g,

$$\begin{aligned} (\nabla^2 \ell_n(\theta'))^{-1} &\xrightarrow{\text{p}} (\nabla^2 \ell_n(\theta_0))^{-1} \\ &\xrightarrow{\text{p}} \left(\sum_{m>0} q(m) \sum_{j=1}^{k} \lambda_j \, \mathsf{E}_{\theta_0}(\nabla^2 V_{jm}) \right)^{-1} \\ &= - \left(\sum_{m>0} q(m) \sum_{j=1}^{k} \lambda_j \, \mathsf{Var}_{\theta_0}(\nabla V_{jm}) \right)^{-1}. \end{aligned} \tag{83}$$

where in the last equality we used a well known identity concerning the Fisher information.

For the remaining term on the rhs of (82) we have

$$\sqrt{n}\,\nabla\ell_n(\theta_0) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} W_i \tag{84}$$

where the random vectors

$$W_i = \sum_{m>0} 1_{\{\text{length}(Y^{(i)})=m\}} \sum_{j=1}^{k} Z_j^{(i)} \nabla V_{jm}^{(i)}$$

have expectation 0 due to the fact that the expectation of the score is 0. The variance of W_i is

$$\begin{aligned}\mathsf{Var}_{\theta_0} W_i &= \mathsf{E}_{\theta_0} \sum_{m>0} 1_{\{\text{length}(Y^{(i)})=m\}} \sum_{j=1}^{k} Z_j^{(i)} \nabla V_{jm}^{(i)} \nabla V_{jm}^{(i)\top} \\ &= \sum_{m>0} q(m) \sum_{j=1}^{k} \lambda_j \,\mathsf{E}\left(\nabla V_{jm}^{(i)} \nabla V_{jm}^{(i)\top}\right)\end{aligned}$$

where in the first equality we used the fact that $Y^{(i)}$ can have only one length and only one of $\chi_1, \ldots, \chi_k$ is chosen. Using the central limit theorem we thus conclude that

$$\sqrt{n}\,\nabla\ell_n(\theta_0) \rightsquigarrow N\left(0, \Sigma\right) \tag{85}$$

and finish the proof by combining (82), (83), and (85) using Slutsky's theorem. □

Figure 25 (left, middle) displays the test-set log-perplexity for the CoNLL2000 chunking task as a function of the total number of labeled tokens. We used the Boltzmann chain MRF model that is the MRF corresponding to HMM (though not identical e.g., [37]). We consider labeling policies $\wp$ that label the entire sequence with probability λ and otherwise label contiguous sequences of length 5 (left) or leave the sequence fully unlabeled (middle). Lighter nodes indicate larger n and unsurprisingly show a decrease in the test-set perplexity as n is increased. Interestingly, the middle figure shows that labeling policies using a smaller amount of labels may outperform other policies. This further motivates our analysis and indicates that naive choices of $\wp$ may inflate labeling cost with negligible improvement to accuracy (cf. Sec. 5.8 for avoiding this pitfall).

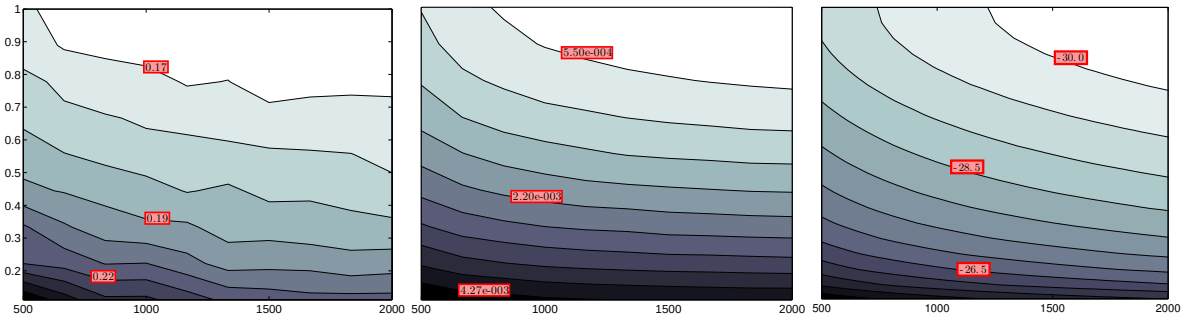

Figure 24: Three error measures for the multinomial naive Bayes SSL classifier applied to Reuters RCV1 text data. In each, error is a function of n (horizontal axis) and λ (vertical axis). The left depicts classification error rate, the middle depicts the trace of empirical MSE, and right depicts the log-trace of the asymptotic variance. Results were obtained using held-out sets and averaged using cross validation. Particularly noteworthy is a striking correlation among all three figures, justifying the use of asymptotic variance as a surrogate for classification error, even for relatively small values of n.

5.7.1 Conditional Structured Prediction

Thus far our discussion on structured prediction has been restricted to generative models such as HMM or Boltzmann chain MRF. Similar techniques, however, can be used to analyze SSL for conditional models such as CRFs that are estimated by maximizing the conditional likelihood. The key to extending the results in this chapter to CRFs is to express conditional SSL estimation in a form similar to (72)

$$\hat{\theta}_n = \arg\max \sum_{i=1}^{n} \log p_\theta(\wp(Y^{(i)})|X^{(i)})$$

and to proceed with an asymptotic analysis that extends the classical conditional MLE asymptotics. We omit further discussion due to lack of space but include some experimental results for CRFs.

Figure 26 (left) depicts a similar experiment to the one described in the previous section for conditional estimation in CRF models. The figure displays log-perplexity as a function of n (x axis) and λ_1 (y axis). We observe a trend nearly identical to that of the Boltzmann chain MRF (Figure 25, left, middle).

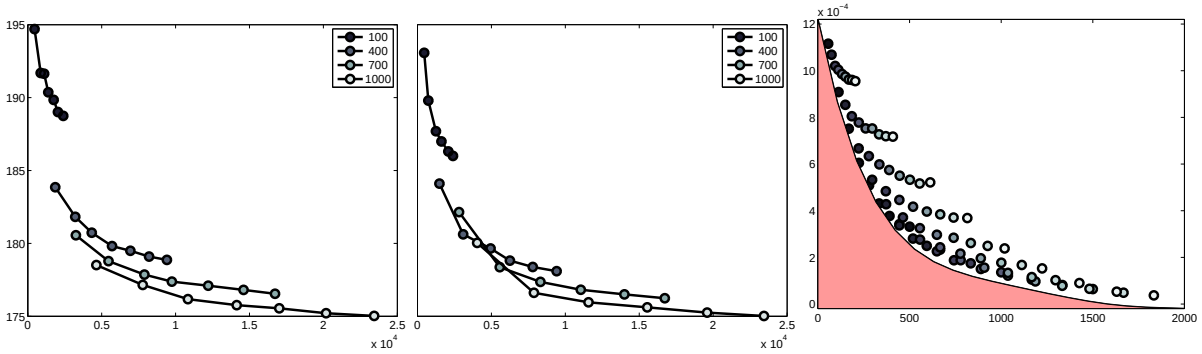

Figure 25: Test-set results for two policies of unlabeled data for Boltzmann chain MRFs applied to the CoNLL 2000 text-chunking dataset (left, middle). The shaded portion of the right panel depicts the empirically unachievable region for naive Bayes SSL classifier on the 20-newsgroups dataset. The left two share a common log-perplexity scale (vertical axis) while the vertical axis of the right panel corresponds to trace of the empirical MSE; the horizontal axis indicates labeling cost. As above, results were obtained using held-out sets and averaged using cross validation. Collectively these figures represent the application and effect of various labeling policies. The left figure depicts the consequence of partially missing samples for various n,λ while the middle and right represent SSL in the more traditional all or nothing sense: either labeled or unlabeled samples. See text for more details.

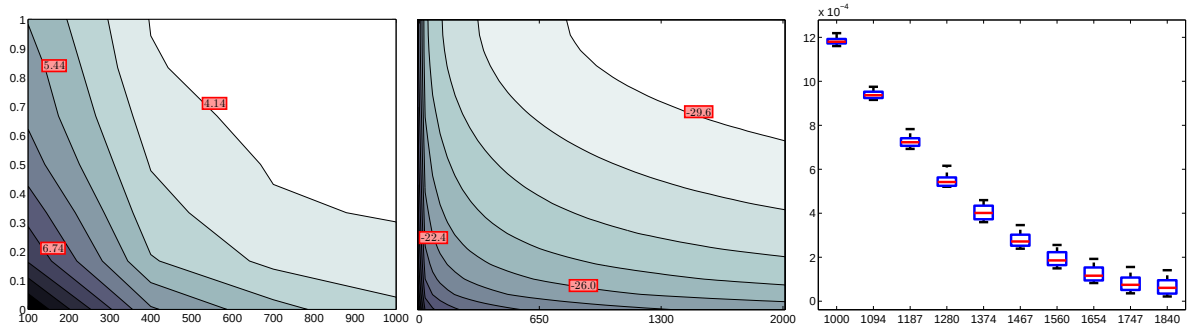

Figure 26: Left figure depicts sentence-wise log-perplexity for CRFs under the same policy and experimental design of the above Boltzmann chain. Center figure represents log-trace of the theoretical variance and demonstrates phenomena under a simplified scenario, i.e., a mixture of two 1000-dim multinomials with unbalanced prior. Rightmost figure demonstrates the practical applicability of utilizing asymptotic analysis to characterize parameter error as a function of size of training-set partition. The training-set is fixed at 2000 samples and split for training and validating. As the proportion used for training is increased, we see a decrease in error. See text for more details.

5.8 A5: Tradeoff

As the figures in the previous sections display, the estimation accuracy increases with the total number of labels. The Cramer-Rao lower bound states that the highest accuracy for unbiased estimators is obtained by the maximum likelihood operating on fully observed data. However, assuming that a certain cost is associated with labeling data SSL resolves a fundamental accuracy-cost tradeoff. A decrease in estimation accuracy is acceptable in return for decreased labeling cost.

Our ability to mathematically characterize the dependency of the estimation accuracy on the labeling cost leads to a new quantitative formulation of this tradeoff. Each labeling policy (λ, n in classification and $\wp$ in structured prediction) is associated with a particular estimation accuracy via Propositions 8 and 10 and with a particular labeling cost. The precise way to measure labeling cost depends on the situation at hand, but we assume in this chapter that the labeling cost is proportional to the number of labeled samples (classification) and of labeled sequence elements (structured prediction). This assumption may be easily relaxed by using other labeling cost functions, e.g, obtaining unlabeled data may incur some cost as well.

Geometrically, each labeling policy may thus be represented in a two dimensional scatter plot where the horizontal and vertical coordinates correspond to labeling cost and estimation error respectively. Three such scatter plots appear in Figure 25 (see Section 5.7 for a description of the left and middle panels). The right panel corresponds to multinomial naive Bayes SSL classifier and the 20-newsgroups classification dataset. Each point in that panel corresponds to different n, λ.

The origin corresponds to the most desirable (albeit unachievable) position in the scatter plot representing zero error at no labeling cost. The cloud of points obtained by varying n, λ (classification) and $\wp$ (structured prediction) represents the achievable region of the diagram. Most attractive is the lower and left boundary of that region which represents labeling policies that dominate others in both accuracy and labeling

cost. The non-achievable region is below and to the left of that boundary (see shaded region in Figure 25, right). The precise position of the optimal policy on the boundary of the achievable region depends on the relative importance of minimizing estimation error and minimizing labeling cost. A policy that is optimal in one context may not be optimal in a different context.

It is interesting to note that even in the case of naive Bayes classification (Figure 25, right) some labeling policies (corresponding to specific choices of n, λ) are suboptimal. These policies correspond to points in the interior of the achievable region. A similar conclusion holds for Boltzmann chain MRF. For example, some of the points in Figure 25 (left) denoted by the label 700 are dominated by the more lightly shaded points.

We consider three different ways to define an optimal labeling policy (i.e., determining how much data to label) on the boundary of the achievable region:

$$(\lambda^*, n^*)_1 = \underset{(\lambda,n):\lambda n \leq C}{\arg\min} \ \mathsf{tr}(\Sigma^{-1}) \tag{86}$$

$$(\lambda^*, n^*)_2 = \underset{(\lambda,n):\mathsf{tr}(\Sigma^{-1}) \leq C}{\arg\min} \ \lambda n \tag{87}$$

$$(\lambda^*, n^*)_3 = \underset{(\lambda,n)}{\arg\min} \ \lambda n + \alpha \ \mathsf{tr}(\Sigma^{-1}). \tag{88}$$

The first applies in situations where the labeling cost is bounded by a certain available budget. The second applies when a certain estimation accuracy is acceptable and the goal is to minimize the labeling cost. The third considers a more symmetric treatment of the estimation accuracy and labeling cost.

Equations (86)-(88) may be easily generalized to arbitrary labeling costs $f(n, \lambda)$. Equations (86)-(88) may also be generalized to the case of structured prediction with $\wp$ replacing (λ, n) and $\text{cost}(\wp)$ replacing λn.

5.9 A6: Practical Algorithms

Choosing a policy (λ, n) or $\wp$ resolves the SSL tradeoff of accuracy vs. cost. Such a resolution is tantamount to answering the basic question of how many labels should be obtained (and in the case of structured prediction also which ones). Resolving the tradeoff via (86)-(88) or in any other way, or even simply evaluating the asymptotic accuracy $\mathsf{tr}(\Sigma)$ requires knowledge of the model parameter θ_0 that is generally unknown in practical settings.

We propose in this section a practical two stage algorithm for computing an estimate $\hat{\theta}_n$ within a particular accuracy-cost tradeoff. Assuming we have n unlabeled examples, the algorithm begins the first stage by labeling r samples. It then estimates θ' by maximizing the likelihood over the r labeled and $n-r$ unlabeled samples. The estimate $\hat{\theta}'$ is then used to obtain a plug-in estimate for the asymptotic accuracy $\mathsf{tr}(\Sigma)$. In the second stage the algorithm uses the estimate $\widehat{\mathsf{tr}(\Sigma)}$ to resolve the tradeoff via (86)-(88) and determine how many more labels should be collected. Note that the labels obtained at the first stage may be used in the second stage as well with no adverse effect.

The two-stage algorithm spends some initial labeling cost in order to obtain an estimate for the quantitative tradeoff parameters. The final labeling cost, however, is determined in a principled way based on the relative importance of accuracy and labeling cost via (86)-(88). The selection of the initial number of labels r is important and should be chosen carefully. In particular it should not exceed the total desirable labeling cost.

We provide some experimental results on the performance of this algorithm in Figure 26 (right). It displays box-plots for the differences between $\mathsf{tr}(\Sigma)$ and $\widehat{\mathsf{tr}(\Sigma)}$ as a function of the initial labeling cost r for naive Bayes SSL classifier and 20-newsgroups data. The figure illustrates that the two stage algorithm provides a very accurate estimation of $\mathsf{tr}(\Sigma)$ for $r \geq 1000$ which becomes almost perfect for $r \geq 1300$.

5.10 Discussion

In this chapter we developed a stochastic m-estimator for controlling labeling costs. Through the SME framework, we able to provide asymptotic analysis of classification and structured prediction tasks under semi-supervised learning scenarios. This analysis allowed us to answer several questions one might reasonably ask in fitting these models. This included addressing how combinations of labeled and unlabeled data lead to precise models, expressing the estimation accuracy as a function of the amount of labeled and unlabeled data, and quantifying the improvement resulting from replacing an unlabeled example with a labeled.

The answers to these questions allowed us to develop practical solutions for budgeted learning scenarios. Such situations are common to many fields, particularly Natural Language Processing, where data evidence is numerous and labels scarce. Through the SME framework we could quantify the value of labels and thereby optimally utilize fiscal resources under several reasonable scenarios. These included fixed budget, minimum required accuracy, and striking a balance between the two.

CHAPTER VI

CONCLUSION

This dissertation aimed to develop, through theory and example, a general mathematical framework for controlling those forces in machine learning which impede accuracy. These limiting factors, which we abstractly referred to as costs, may be computational or they may be financial.

In chapter 2 we began construction of such a framework by considering the broad class of estimators known as m-estimators. Such estimators consist of maximizing an average and are ubiquitous in statistics and machine learning. Notable special cases include least-squares estimation and the maximum likelihood estimator. We improved upon the standard m-estimator by imagining a "super" m-estimator, comprised of selecting several criteria functions at-random, and maximizing their sum. We called this the stochastic m-estimator (SME) and were able to show that each SME instantiation resolves the accuracy-cost tradeoff differently, and taken together they span a continuous spectrum of accuracy-cost tradeoff resolutions.

We also proved that the stochastic m-estimator inherits the desirable properties of its m-estimator predecessor, including consistency and asymptotic Normality. Such properties are fairly modest, owing to their asymptotic nature, but nonetheless characterize the estimator in an intuitive and fundamentally significant manner. Simply put, satisfying these properties ensures that given sufficiently many data, the estimator will behave in a manner which is faithful to the distribution which underlies the data.

In chapters 2, 3, and 4 we applied the SME to several fundamental tasks in machine learning. In chapter 2 we explored the computational challenges of learning

parameters in broad class of models known as Markov random fields (MRFs). MRFs include simple models, such as the Exponential and Poisson distributions, as well as more sophisticated models, such the Boltzmann machine and the Ising model. In chapter 3 we continued to study computational tradeoffs in MRF parameter learning, only this time tailored the SME to handle models with latent random variables. We called this extension the MCEM+SCL hybrid. Notable latent variable MRFs include the Mixture of Gaussians and Latent Dirichlet Allocation [8]. Through theoretical and experimental study, we demonstrated the effectiveness of the estimators when computational resources are insufficient or when obtaining additional labeled samples is necessary. We also demonstrated that in some cases the stochastic m-estimator is associated with robustness thereby increasing its statistical accuracy and representing a win-win.

In chapter 4, we examined another cost which is common to every machine learning problem: label complexity. That is, how do we obtain adequately many training samples to ensure accurate inference. We asked, and answered, several questions pertaining to a particular means of tolerating partially labeled samples known as semi-supervised learning. As a happy side-effect, we discovered that the stochastic m-estimator framework yields some intuition as to how much a particular partial labeling will improve estimator accuracy. From this idea we developed a simple procedure for requesting more labeled samples and in what configuration they should be labeled.

We imagine the SME framework as being appropriate for several other tasks in statistics and machine learning. It may be possible to extend ideas similar to stochastic m-estimation to sample generation, much like blocked Gibbs sampling. It would also seem that feature learning is a viable candidate for the SME. One may imagine the features themselves as being randomly selected and the SME framework could be used to characterize the performance of different selection policies. This idea is

important when there are too many features to collect, perhaps due to insufficient computer memory, or when there are varying costs associated with different features.

As an illustrative example, consider the task of information retrieval and the aggressive time tradeoff the modern search engine must negotiate. Typically this task involves ranking an enormous collection of documents in some short interval, say 250 milliseconds, concurrently for hundreds of thousands of users. Document relevance is based not only on quality and similarity to the query, but also on characteristics specific to the user. Such characteristics or features may include the user's demographic (gender, age, location, interests), past session data, or features based on other similar users (users of a type clicked on pages of a type). Each of these features are likely to require increasing amounts of computation and network time, but they are also likely to improve retrieval quality. We believe the SME is a natural candidate to describe this tradeoff. It offers a methodology for developing approaches which balance the cost of obtaining richer features and the improvement those features bring to user satisfaction.

As this dissertation demonstrated, we believe the stochastic m-estimator framework is robust enough to rise to many of the accuracy-cost tradeoffs present in machine learning. We also hope that its fundamentally simple construction makes it a compelling tool in the machine learning practitioner's arsenal.

APPENDIX A

PROOFS

A.1 m*-Estimator*

Theorem 8 (m-Estimator Strong Consistency)**.** *The following assumptions imply that the sequence of estimators $\hat{\theta}_n$ is strongly consistent with some $\theta_0 \in \Theta_0$, i.e., there exists $\theta_0 \in \Theta_0$ such that,*

$$p_0 \left\{ \lim_{n\to\infty} d(\hat{\theta}_n, \theta_0) = 0 \right\} = 1. \tag{89}$$

Assumptions:

(A1) *The set Θ is compact and permits metric $d(\cdot,\cdot)$.*

(A2) *The function $m : \Theta \times \mathcal{X} \to \{-\infty, \infty\} \cup \mathbb{R}$ is upper semi-continuous in θ for almost all $x \in \mathcal{X}$, i.e., for all $\theta \in \Theta$,*

$$p_0 \left\{ \limsup_{\theta'\to\theta} m_\theta(X) \leq m_\theta(X) \right\} = 1.$$

(A3) *Write $\psi(x,\theta,\rho) = \sup\{m_{\theta'}(x) : d(\theta',\theta) < \rho\}$. For every $\theta \in \Theta$ and sufficiently small $\rho > 0$:*

(i) *The function $\psi(x,\theta,\rho)$ is ν-measurable and,*

(ii) *The population expectation is positive integrable, i.e., $\mathsf{E}_0\,\psi(X,\theta,\rho) < \infty$.*

(A4) *The estimator $\hat{\theta}_n$ nearly maximizes M_n almost everywhere, i.e., $p_0(M_n(\hat{\theta}_n) \geq M_n(\theta_0)) = 1$.*

Proof.

If Θ is empty or $M_0(\theta)$ is identically $-\infty$ then $\Theta_0 = \Theta$ and the theorem is vacuous, hence we may assume $\Theta_0 \subset \Theta \neq \emptyset$. Since Θ is compact (A1), the supremum are achieved on Θ and $\Theta_0 \neq \emptyset$, so $\mathsf{E}_0\, m_{\theta_0} > -\infty$. This fact together with A3(ii) implies $\mathsf{E}_0\, |m_{\theta_0}| < \infty$ so $|M_0(\theta_0)| < \infty$.

(We next show limits and integrals are exchangeable.)

Let $\rho \geq \rho_n$ be a decreasing sequence in n. Fix some θ; the sequence $\psi(x, \theta, \rho_n)$ is decreasing and bounded above by $\psi(x, \theta, \rho)$. In view of A2 and $m_\theta(x) \leq \psi(x, \theta, \rho_n)$ we have $\psi(x, \theta, \rho_n) \downarrow m_\theta(x)$ as $n \to \infty$. By A3, $\psi(x, \theta, \rho)$ is measurable and positive integrable. These facts satisfy the conditions of the (extended) monotone convergence theorem[1] and we conclude $\mathsf{E}_0\, \psi(X, \theta, \rho_n) \downarrow M_0(\theta) \geq -\infty$ as $n \to \infty$.

(We next establish a finite subcover of a subset of Θ which excludes open neighborhoods Θ_0.)

In view of the preceding paragraph and the fact that $M_0(\theta) < M_0(\theta_0)$ for every $\theta \notin \Theta_0$ we have that for every θ there exists $\rho_\theta > 0$ such that $\mathsf{E}_0\, \psi(X, \theta, \rho_\theta) < M_0(\theta_0)$. Write $B_\theta = B(\theta, \rho_\theta)$.

Let $S_\varepsilon = \bigcap_{\theta_0 \in \Theta_0} B^c(\theta, \varepsilon)$ for some $\varepsilon > 0$. The set S_ε is covered by $\{B_\theta : \theta \in S_\varepsilon\}$, i.e., $S_\varepsilon \subset \cup_{\theta \in S_\varepsilon} B_\theta$. Note that the cover excludes members of Θ_0 by definition of B_θ. Since S_ε is an intersection of closed sets it is compact. Let $\{B_{\theta_j} : j = 1, \ldots, p\}$ be a finite subcover of S_ε.

(We now invoke the strong law of large numbers on the finite sub-cover.)

Write $\mathsf{E}_n\, \psi_j = \mathsf{E}_n\, \psi(X, \theta_j, \rho_{\theta_j})$. We apply the SLLN[2] to each of the finitely many subcover elements and conclude,

$$p_0 \left\{ \lim_{n \to \infty} \mathsf{E}_n\, \psi(X, \theta_j, \rho_j) = \mathsf{E}_0\, \psi(X, \theta_j, \rho_j) \right\} = 1,$$

for each $j = 1, \ldots, p$. Since the finite intersection of almost sure events remains

[1] Ash & Doléans-Dade. Page 49. Theorem 1.6.7(b).
[2] Ash & Doléans-Dade. Pages 242–244. Theorem 6.2.5.

almost sure we have,

$$\sup_{\theta \in S_\varepsilon} M_n(\theta) \leq \max_{1 \leq j \leq p} \mathsf{E}_n \, \psi_j \xrightarrow{\text{as}} \max_{1 \leq j \leq p} \mathsf{E}_0 \, \psi_j < M_0(\theta_0),$$

and conclude,

$$p_0 \left\{ \lim_{n \to \infty} \sup_{\theta \in S_\varepsilon} M_n(\theta) < M_0(\theta_0) \right\} = 1. \tag{90}$$

(We conclude the proof by showing $\hat{\theta}_n \notin S_\varepsilon$ by assuming its complement.)

If we assume $\hat{\theta}_n \in S_\varepsilon$ then obviously $M_n(\hat{\theta}_n) \leq \sup_{\theta \in S_\varepsilon} M_n(\theta)$. This fact, A4, and the pointwise application of the SLLN implies that,

$$\sup_{\theta \in S_\varepsilon} M_n(\theta) \geq M_n(\hat{\theta}_n) \geq M_n(\theta_0) \xrightarrow{\text{as}} M_0(\theta_0),$$

and the conclusion,

$$p_0 \left\{ M_0(\theta) \leq \lim_{n \to \infty} \sup_{\theta \in S_\varepsilon} M_n(\theta) \right\} = 1. \tag{91}$$

However, combining (90) and (91) yields the contradiction,

$$\begin{aligned} & p_0 \left\{ M_0(\theta_0) \leq \lim_{n \to \infty} \sup_{\theta \in S_\varepsilon} M_n(\theta) < M_0(\theta_0) \right\} = 1 \\ & = p_0 \left\{ M_0(\theta_0) < M_0(\theta_0) \right\} \end{aligned}$$

which implies $\hat{\theta}_n \notin S_\varepsilon$. Since ε was chosen arbitrarily the theorem follows. This proof is due to Wald. □

A.2 Stochastic Composite Likelihood

The proofs below generalize the classical consistency and asymptotic efficiency of the MLE [24] and the corresponding results for m-estimators [49]. They follow similar lines as the proofs in [24] and [49], with the necessary modifications due to the stochasticity of the SCL function. We assume below that $p_\theta(X) > 0$ and that X is a discrete and finite RV.

The following lemma generalizes Shannon's inequality [17] for the KL divergence. We will use it to prove consistency of the SCL estimator.

Lemma A.2.1. Let $(A_1, B_1), \ldots, (A_k, B_k)$ be a sequence of m-pairs that ensures identifiability of $p_\theta, \theta \in \Theta$ and $\alpha_1, \ldots, \alpha_k$ positive constants. Then

$$\sum_{j=1}^{k} \alpha_k \, D(p_\theta(X_{A_j}|X_{B_j}) \,||\, p_{\theta'}(X_{A_j}|X_{B_j})) \geq 0 \tag{92}$$

where equality holds iff $\theta = \theta'$.

Proof. The inequality follows from applying Jensen's inequality for each conditional KL divergence

$$\begin{aligned} -D(p_\theta(X_{A_j}|X_{B_j}) \,||\, p_{\theta'}(X_{A_j}|X_{B_j})) = \mathsf{E}_{p_\theta} \log \frac{p_{\theta'}(X_{A_j}|X_{B_j})}{p_\theta(X_{A_j}|X_{B_j})} &\leq \log E_{p_\theta} \frac{p_{\theta'}(X_{A_j}|X_{B_j})}{p_\theta(X_{A_j}|X_{B_j})} \\ &= \log 1 = 0. \end{aligned}$$

For equality to hold we need each term to be 0 which follows only if $p_\theta(X_{A_j}|X_{B_j}) \equiv p_{\theta'}(X_{A_j}|X_{B_j})$ for all j which, assuming identifiability, holds iff $\theta = \theta'$. □

Proposition 1. Let $\Theta \subset \mathbb{R}^r$ be an open set, $p_\theta(x) > 0$ and continuous and smooth in θ, and $(A_1, B_1), \ldots, (A_k, B_k)$ be a sequence of m-pairs for which $\{(A_j, B_j) : \forall j \text{ such that } \lambda_j > 0\}$ ensures identifiability. Then the sequence of SCL maximizers is strongly consistent, i.e.,

$$P\left(\lim_{n\to\infty} \hat{\theta}_n = \theta_0\right) = 1. \tag{93}$$

Proof. The SCL function, modified slightly by a linear combination with a term that is constant in θ is

$$sc\ell'(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \beta_j \left(Z_{ij} \log p_\theta(X_{A_j}^{(i)}|X_{B_j}^{(i)}) - \lambda_j \log p_{\theta_0}(X_{A_j}^{(i)}|X_{B_j}^{(i)}) \right).$$

By the strong law of large numbers, the above expression converges as $n \to \infty$ to its expectation

$$\mu(\theta) = -\sum_{j=1}^{k} \beta_j \lambda_j \, D(p_{\theta_0}(X_{A_j}|X_{B_j}) \,||\, p_\theta(X_{A_j}|X_{B_j})).$$

If we restrict ourselves to the compact set $S = \{\theta : c_1 \leq \|\theta - \theta_0\| \leq c_2\}$ then

$$\sup_{\theta \in S} \sup_{Z} \left| \sum_{j=1}^{k} Z_j \beta_j \log p_\theta(X_{A_j}|X_{B_j}) - \lambda_j \beta_j \log p_{\theta_0}(X_{A_j}|X_{B_j}) \right| < K(x) < \infty \tag{94}$$

where $K(x)$ is a function satisfying $\mathsf{E}\, K(X) < \infty$. As a result, the conditions for the uniform strong law of large numbers [24] hold on S leading to

$$P\left\{ \lim_{n \to \infty} \sup_{\theta \in S} |scl'(\theta) - \mu(\theta)| = 0 \right\} = 1. \tag{95}$$

By Proposition A.2.1, $\mu(\theta)$ is non-positive and is zero iff $\theta = \theta_0$. Since the function $\mu(\theta)$ is continuous it attains its negative supremum on the compact S: $\sup_{\theta \in S} \mu(\theta) < 0$. Combining this fact with (95) we have that there exists N such that for all $n > N$ the SCL maximizers on S achieves strictly negative values of $sc\ell'(\theta)$ with probability 1. However, since $sc\ell'(\theta)$ can be made to achieve values arbitrarily close to zero under $\theta = \theta_0$, we have that $\hat{\theta}_n^{\text{msl}} \notin S$ for $n > N$. Since c_1, c_2 were chosen arbitrarily $\hat{\theta}_n^{\text{msl}} \to \theta_0$ with probability 1. □

Proposition 2. Making the assumptions of Proposition 1 as well as convexity of $\Theta \subset \mathbb{R}^r$ we have the following convergence in distribution

$$\sqrt{n}(\hat{\theta}_n^{\text{msl}} - \theta_0) \rightsquigarrow N\left(0, \Upsilon \Sigma \Upsilon\right) \tag{96}$$

where

$$\Upsilon^{-1} = \sum_{j=1}^{k} \beta_j \lambda_j \, \mathsf{Var}_{\theta_0}(\nabla S_{\theta_0}(A_j, B_j)) \tag{97}$$

$$\Sigma = \mathsf{Var}_{\theta_0}\left(\sum_{j=1}^{k} \beta_j \lambda_j \nabla S_{\theta_0}(A_j, B_j) \right). \tag{98}$$

The notation $\mathsf{Var}_{\theta_0}(Y)$ represents the covariance matrix of the random vector Y under p_{θ_0} while the notations $\xrightarrow{\text{p}}, \rightsquigarrow$ in the proof below denote convergences in probability and in distribution [24].

Proof. By the mean value theorem and convexity of Θ there exists $\eta \in (0,1)$ for which $\theta' = \theta_0 + \eta(\hat{\theta}_n^{\text{msl}} - \theta_0)$ and

$$\nabla sc\ell_n(\hat{\theta}_n^{\text{msl}}) = \nabla sc\ell_n(\theta_0) + \nabla^2 sc\ell_n(\theta')(\hat{\theta}_n^{\text{msl}} - \theta_0)$$

where $\nabla f(\theta)$ and $\nabla^2 f(\theta)$ are the $r \times 1$ gradient vector and $r \times r$ matrix of second order derivatives of $f(\theta)$. Since $\hat{\theta}_n$ maximizes the SCL, $\nabla sc\ell_n(\hat{\theta}_n^{\text{msl}}) = 0$ and

$$\sqrt{n}(\hat{\theta}_n^{\text{msl}} - \theta_0) = -\sqrt{n}(\nabla^2 sc\ell_n(\theta'))^{-1} \nabla sc\ell_n(\theta_0). \tag{99}$$

By Proposition 1 we have $\hat{\theta}_n^{\text{msl}} \xrightarrow{\text{p}} \theta_0$ which implies that $\theta' \xrightarrow{\text{p}} \theta_0$ as well. Furthermore, by the law of large numbers and the fact that if $W_n \xrightarrow{\text{p}} W$ then $g(W_n) \xrightarrow{\text{p}} g(W)$ for continuous g,

$$\begin{aligned}(\nabla^2 sc\ell_n(\theta'))^{-1} &\xrightarrow{\text{p}} (\nabla^2 sc\ell_n(\theta_0))^{-1} \\ &\xrightarrow{\text{p}} \left(\sum_{j=1}^{k} \beta_j \lambda_j \, \mathsf{E}_{\theta_0} \, \nabla^2 S_{\theta_0}(A_j, B_j) \right)^{-1} \\ &= - \left(\sum_{j=1}^{k} \beta_j \lambda_j \, \mathsf{Var}_{\theta_0}(\nabla S_{\theta_0}(A_j, B_j)) \right)^{-1}.\end{aligned} \tag{100}$$

For the remaining term in (99) we have

$$\sqrt{n} \, \nabla sc\ell_n(\theta_0) = \sum_{j=1}^{k} \beta_j \sqrt{n} \, \frac{1}{n} \sum_{i=1}^{n} W_{ij}$$

where the random vectors $W_{ij} = Z_{ij} \nabla \log p_\theta(X_{A_j}^{(i)} | X_{B_j}^{(i)})$ have expectation 0 and variance matrix $\mathsf{Var}_{\theta_0}(W_{ij}) = \lambda_j \, \mathsf{Var}_{\theta_0}(\nabla S_{\theta_0}(A_j, B_j))$. By the central limit theorem

$$\sqrt{n} \, \frac{1}{n} \sum_{i=1}^{n} W_{ij} \rightsquigarrow N\left(0, \lambda_j \, \mathsf{Var}_{\theta_0}(\nabla S_{\theta_0}(A_j, B_j))\right).$$

The sum $\sqrt{n} \, \nabla sc\ell_n(\theta_0) = \sum_{j=1}^{k} \beta_j \sqrt{n} \, \frac{1}{n} \sum_{i=1}^{n} W_{ij}$ is asymptotically Gaussian as well with mean zero since it converges to a sum of Gaussian distributions with mean zero. Since in the general case the random variables $\sqrt{n} \, \frac{1}{n} \sum_{i=1}^{n} W_{ij}$, $j = 1, \ldots, k$

are correlated, the asymptotic variance matrix of $\sqrt{n}\,\nabla sc\ell_n(\theta_0)$ needs to account for cross covariance terms leading to

$$\sqrt{n}\,\nabla sc\ell_n(\theta_0) \rightsquigarrow N\left(0, \mathsf{Var}_{\theta_0}\left(\sum_{j=1}^{k} \beta_j \lambda_j \nabla S_{\theta_0}(A_j, B_j)\right)\right). \tag{101}$$

We finish the proof by combining (99), (100) and (101) using Slutsky's theorem. □

Recall our notation for the case that the true model $P \notin \{p_\theta : \theta \in \Theta\}$.

$$\psi_\theta(X, Z) \stackrel{\text{def}}{=} \nabla m_\theta(X, Z) \tag{102}$$

$$\dot{\psi}_\theta(X, Z) \stackrel{\text{def}}{=} \nabla^2 m_\theta(X, Z) \quad \text{(matrix of second order derivatives)} \tag{103}$$

$$\Psi_n(\theta) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X^{(i)}, Z^{(i)}). \tag{104}$$

Proposition 3. Assuming the conditions in Proposition 1 as well as $\sup_{\theta:\|\theta-\theta_0\|\geq\epsilon} M(\theta) < M(\theta_0)$ for all $\epsilon > 0$ we have $\hat{\theta}_n^{\text{msl}} \to \theta_0$ as $n \to \infty$ with probability 1.

Proof. We assert

$$P\left\{\lim_{n\to\infty} \sup_{\theta\in S} |scl'(\theta) - \mu(\theta)| = 0\right\} = 1. \tag{105}$$

on the compact set $S = \{\theta : c_1 \leq \|\theta - \theta_0\| \leq c_2\}$ as in the proof of Proposition 1. We proceed similarly along the lines of Proposition 1, with the necessary modification due to the fact that the true model is outside the parametric family.

Since the function $\mu(\theta)$ is continuous it attains its negative supremum on the compact S: $\sup_{\theta\in S} \mu(\theta) < \mu(\theta_0) \geq 0$. Combining this fact with (105) we have that there exists N such that for all $n > N$ the SCL maximizers on S achieves strictly negative values of $sc\ell'(\theta)$ with probability 1.

However, since $sc\ell'(\theta)$ can be made to achieve values arbitrarily close to $\mu(\theta_0)$ as $\hat{\theta}_n \to \theta_0$, we have that $\hat{\theta}_n^{\text{msl}} \notin S$ for $n > N$. Since c_1, c_2 were chosen arbitrarily $\hat{\theta}_n^{\text{msl}} \to \theta_0$ with probability 1. □

Proposition 4. Assuming the conditions of Proposition 2 as well as $\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\|\psi_{\theta_0}(X,Z)\|^2 < \infty$, $\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\dot{\psi}_{\theta_0}(X)$ exists and is non-singular, $|\ddot{\Psi}_{ij}| = |\partial^2\psi_\theta(x)/\partial\theta_i\theta_j| < g(x)$ for all i,j and θ in a neighborhood of θ_0 for some integrable g, we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -(\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\dot{\psi}_{\theta_0})^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\theta_0}(X^{(i)},Z^{(i)}) + o_P(1) \tag{106}$$

or equivalently

$$\hat{\theta}_n = \theta_0 - (\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\dot{\psi}_{\theta_0})^{-1}\frac{1}{n}\sum_{i=1}^{n}\psi_{\theta_0}(X^{(i)},Z^{(i)}) + o_P\left(\frac{1}{\sqrt{n}}\right). \tag{107}$$

Proof. By Taylor's theorem there exists a random vector $\tilde{\theta}_n$ on the line segment between θ_0 and $\hat{\theta}_n$ for which

$$0 = \Psi_n(\hat{\theta}_n) = \Psi_n(\theta_0) + \dot{\Psi}_n(\theta_0)(\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^\top\ddot{\Psi}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0).$$

which we re-arrange as

$$\sqrt{n}\dot{\Psi}_n(\theta_0)(\hat{\theta}_n - \theta_0) + \sqrt{n}\frac{1}{2}(\hat{\theta}_n - \theta_0)^\top\ddot{\Psi}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0) = -\sqrt{n}\Psi_n(\hat{\theta}_n) \tag{108}$$

$$= -\sqrt{n}\Psi_n(\theta_0) + o_P(1) \tag{109}$$

where the second equality follows from the fact that $\hat{\theta}_n \xrightarrow{\text{P}} \theta_0$ and continuous functions preserves converges in probability.

Since $\dot{\Psi}_n(\theta_0)$ converges by the law of large numbers to $\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\dot{\psi}_\theta(X,Z)$ and $\ddot{\Psi}_n(\tilde{\theta}_n)$ converges to a matrix of bounded values in the neighborhood of θ_0 (for large n), the lhs of (108) is

$$\sqrt{n}\left(\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\dot{\psi}_\theta(X,Z) + o_P(1) + \frac{1}{2}(\hat{\theta}_n - \theta_0)O_P(1)\right)(\hat{\theta}_n - \theta_0)$$

$$= \sqrt{n}(\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\dot{\psi}_\theta(X,Z) + o_P(1))(\hat{\theta}_n - \theta_0) \tag{110}$$

since $\hat{\theta}_n - \theta_0 = o_P(1)$ and $o_P(1)O_p(1) = o_P(1)$ (the notation $O_P(1)$ denotes stochastically bounded and it applies to $\ddot{\Psi}_n(\tilde{\theta}_n)$ as described above). Putting it together we

have

$$\sqrt{n}(\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\dot{\psi}_\theta(X,Z)+o_P(1))(\hat{\theta}_n-\theta_0)=-\sqrt{n}\Psi_n(\theta_0)+o_P(1).$$

Since the matrix $\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\dot{\psi}_\theta(X,Z)+o_P(1)$ converges to a non-singular matrix, multiplying the equation above by its inverse finishes the proof. □

Corollary 1. Assuming the conditions specified in Proposition 5 we have

$$\sqrt{n}(\hat{\theta}_n-\theta_0)\rightsquigarrow N(0,(\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\dot{\psi}_{\theta_0})^{-1}(\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\psi_{\theta_0}\psi_{\theta_0}^\top)(\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\dot{\psi}_{\theta_0})^{-1}). \tag{111}$$

Proof. Equation (34) follows from (32) by noticing that due to the central limit theorem $\Psi_n(\theta_0)$ (as it is an average of n iid RVs with expectation 0)

$$\sqrt{n}\cdot\frac{1}{n}\sum_{i=1}^{n}\psi_{\theta_0}(X^{(i)},Z^{(i)})\rightsquigarrow N(0,\mathsf{E}_{P(X)}\,\mathsf{E}_{P(Z)}\,\psi_{\theta_0}\psi_{\theta_0}^\top).$$

Substituting this in the right hand side of (32) and accounting for the modified variance due to the matrix inverse results in (34). □

REFERENCES

[1] Arnold, B. and Strauss, D., "Pseudolikelihood estimation: some examples," *Sankhya B*, vol. 53, pp. 233–243, 1991.

[2] Arnold, B. C., Castillo, E., and Sarabia, J.-M., *Conditional Specification of Statistical Models*. Springer, 1999.

[3] Ash, R. A. and Doleans-Dade, C. A., *Probability and Measure Theory*. Academic Press, second ed., 1999.

[4] Balcan, M. F. and Blum, A., "A discriminative model for semi-supervised learning," *Journal of the Association for Computing Machinery*, 2010.

[5] Ben-David, S., Lu, T., and Pal, D., "Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning," in *International Conference on Learning Theory*, 2008.

[6] Besag, J., "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *Journal of the Royal Statistical Society B*, vol. 36, no. 2, pp. 192–236, 1974.

[7] Bishop, Y., Fienberg, S., and Holland, P., *Discrete multivariate analysis: theory and practice*. MIT press, 1975.

[8] Blei, D., Ng, A., and Jordan, M., "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[9] Bolthausen, E., van der Vaart, A., and Perkins, E., "Lecture: Empirical processes and consistency of z-estimators," in *Lectures on Probability Theory and Statistics*, 2002.

[10] Bottou, L. and Bousquet, O., "Learning using large datasets," in *Mining Massive DataSets for Security*, IOS Press, 2008.

[11] Brook, D., "On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems," in *Biometrika*, 1964.

[12] Cappé, O., Moulines, E., and Rydén, T., *Inference in Hidden Markov Models*. Springer, 2005.

[13] Casella, R. and Robert, C., *Monte Carlo Statistical Methods*. Springer Verlag, second ed., 2004.

[14] Castelli, V. and Cover, T. M., "The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 2102–2117, 1996.

[15] Chapelle, O., Schölkopf, B., and Zien, A., eds., *Semi-Supervised Learning*. MIT Press, 2006.

[16] Cohen, I. and Cozman, F. G., "Risks of semi-supervised learning," in *Semi-Supervised Learning*, MIT press, 2006.

[17] Cover, T. M. and Thomas, J. A., *Elements of Information Theory*. John Wiley & Sons, second ed., 2005.

[18] Cox, D. R. and Snell, E. J., "A general definition of residuals (with discussion)," *Journal of the Royal Statistical Society B*, 1968.

[19] Csiszár, I., "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *The Annals of Statistics*, vol. 19, no. 4, 1991.

[20] Dempster, A. P., Laird, N. M., and Rubin, D. B., "Maximum likelihood from incomplete data via the em algorithm," in *Journal of the Royal Statistical Society. Series B (Methodological).*, 1977.

[21] Dillon, J. and Lebanon, G., "Statistical and computational tradeoffs in stochastic composite likelihood," in *Proc. of the International Conference on Artificial Intelligence and Statistics*, 2009.

[22] Dillon, J. V., Balasubramanian, K., and Lebanon, G., "Asymptotic analysis of generative semi-supervised learning," in *Proc. of the International Conference on Machine Learning*, 2010.

[23] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G., *Biological Sequence Analysis*. Cambridge University Press, 1998.

[24] Ferguson, T. S., *A Course in Large Sample Theory*. Chapman & Hall, 1996.

[25] Fort, G. and Moulines, E., "Convergence of the monte carlo expectation maximization for curved exponential families," in *The Annals of Statistics*, 2003.

[26] Gelman, A. and Speed, T. P., "Characterizing a joint probability distribution by conditionals," in *Journal of the Royal Statistical Society. Series B (Methodological)*, 1993.

[27] Hinton, G. and Sejnowski, T., "Optimal perceptual inference," in *Proc. Computer Vision and Pattern Recognition*, 1983.

[28] Hjort, N. and Varin, C., "ML, PL, and QL in markov chain models," *Scandinavian Journal of Statistics*, vol. 35, no. 1, pp. 64–82, 2008.

[29] Horn, R. and Johnson, C. R., *Matrix Analysis*. Cambridge University Press, 1990.

[30] Hunter, D. R. and Lange, K., "A tutorial on mm algorithms," in *The American Statistician*, 2004.

[31] Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K., "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.

[32] Koller, D. and Friedman, N., *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[33] Lewis, D., Yang, Y., Rose, T., and Li, F., "RCV1: A new benchmark collection for text categorization research," *Journal of Machine Learning Research*, vol. 5, pp. 361–397, 2004.

[34] Liang, G. and Yu, B., "Maximum pseudo likelihood estimation in network tomography," *IEEE Trans. Signal Process*, vol. 51, no. 8, pp. 2043–2053, 2003.

[35] Liang, P. and Jordan, M. I., "An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators.," in *Proc. of the International Conference on Machine Learning*, 2008.

[36] Lindsay, B. G., "Composite likelihood methods," *Contemporary Mathematics*, vol. 80, pp. 221–239, 1988.

[37] MacKay, D. J. C., "Equivalence of linear boltzmann chains and hidden markov models," *Neural Computation*, vol. 8, no. 1, pp. 178–181, 1996.

[38] Mao, Y. and Lebanon, G., "Isotonic conditional random fields and local sentiment flow," in *Advances in Neural Information Processing Systems 19*, pp. 961–968, 2007.

[39] Marlin, B., Swersky, K., Chen, B., and de Freitas, N., "Inductive principles for restricted boltzmann machine learning," *Proc. of the International Conference on Artificial Intelligence and Statistics*, 2010.

[40] Nigam, K., McCallum, A., Thrun, S., and Mitchell, T., "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, vol. 39, no. 2, pp. 103–134, 2000.

[41] Pollard, D., *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, 2001.

[42] Robbins, H. and Monro, S., "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, pp. 400–407, 1951.

[43] Robert, C. and Casella, G., *Monte Carlo Statistical Methods*. Springer, 2004.

[44] Serfling, R. J., *Approximation Theorems of Mathematical Statistics*. John Wiley, 1980.

[45] Sha, F. and Pereira, F., "Shallow parsing with conditional random fields," *Proceedings of HLT-NAACL*, pp. 213–220, 2003.

[46] Singh, A., Nowak, R., and Zhu, X., "Unlabeled data: Now it helps, now it doesnt," in *Advances in Neural Information Processing Systems*, 2008.

[47] Sinha, K. and Belkin, M., "The value of labeled and unlabeled examples when the model is imperfect," in *Advances in Neural Information Processing Systems*, 2008.

[48] Sutton, C. and McCallum, A., "Piecewise pseudolikelihood for efficient training of conditional random fields," in *Proc. of the International Conference on Machine Learning*, 2007.

[49] van der Vaart, A. W., *Asymptotic Statistics*. Cambridge University Press, 1998.

[50] van der Vaart, A. and Wellner, J., *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

[51] Varin, C. and Vidoni, P., "A note on composite likelihood inference and model selection," *Biometrika*, vol. 92, pp. 519–528, 2005.

[52] Vickrey, D., Lin, C., and Koller, D., "Non-local contrastive objectives," in *Proc. of the International Conference on Machine Learning*, 2010.

[53] Wainwright, M. J. and Jordan, M. I., "Graphical models, exponential families, and variational inference," Tech. Rep. 649, UC Berkeley Statistics Department, 2003.

[54] Wainwright, M. J. and Jordan, M. I., "A variational principle for graphical models," in *New Directions in Statistical Signal Processing: From Systems to Brain* (Haykin, S., Principe, J. C., Sejnowski, T. J., and McWhirter, J., eds.), MIT Press, 2005.

[55] Wang, B. and Titterington, D. M., "Lack of consistency of mean field and variational break bayes approximations for state space models," *Neural Processing Letters*, 2004.

[56] Wei, G. and Tanner, M., "A monte-carlo implementation of the em algorithm and the poor man's data augmentation algorithms," in *Journal of the American Statistical Association*, 1991.

[57] Wu, C. F. J., "On the convergence properties of the em algorithm," in *The Annals of Statistics*, 1983.

[58] Xing, E. P., Jordan, M. I., and Russell, S., "A generalized mean field algorithm for variational inference in exponential families," in *Proc. of Uncertainty in Artificial Intelligence*, 2003.

[59] Zhang, T., "The value of unlabeled data for classification problems," in *Proc. of the International Conference on Machine Learning*, 2000.

[60] Zhu, S.-C. and Liu, X., "Learning in Gibbsian fields: How accurate and how fast can it be?," *IEEE Trans. Pattern Analysis*, vol. 24, no. 7, pp. 1001–1006, 2002.